TR-EE65-9
AFCRL-65-681

PURDUE UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

SPEECH ANALYSIS

George W. Hughes John F. Hemdal
Purdue Research Foundation
Lafayette, Indiana

AF 19(628)-305
Project 5628
Task 562802

FINAL REPORT
July 1, 1965
Period Covered: January 1962 – December 1964

PREPARED FOR
AIR FORCE CAMBRIDGE RESEARCH LABORATORIES
OFFICE OF AEROSPACE RESEARCH
UNITED STATES AIR FORCE
BEDFORD, MASSACHUSETTS

Requests for additional copies by agencies of the Department of Defense, their contractors, or other government agancies should be directed to:

DEFENSE DOCUMENTATION CENTER (DDC)
CAMERON STATION
ALEXANDRIA, VIRGINIA 22314

Department of Defense contractors must be established for DDC services or have their "need-to-know" certified by the cognizant military agency of their project or contract.

All other persons and organizations should apply to the:

CLEARINGHOUSE FOR FEDERAL SCIENTIFIC AND
TECHNICAL INFORMATION (CFSTI)
SILLS BUILDING
5285 PORT ROYAL ROAD
SPRINGFIELD, VIRGINIA 22151

AFCRL-65-681 TR EE 65-9

SPEECH ANALYSIS

George W. Hughes John F. Hemdal

Purdue Research Foundation

Lafayette, Indiana

AF 19(628)-305

Project 5628

Task 562802

FINAL REPORT

July 1, 1965

Period Covered: January 1962 - December 1964

Prepared

for

AIR FORCE CAMBRIDGE RESEARCH LABORATORIES

OFFICE OF AEROSPACE RESEARCH

UNITED STATES AIR FORCE

BEDFORD, MASSACHUSETTS

FOREWORD

This report summarizes work done in the Purdue University School of Electrical Engineering under Contract AF 19 628-305 with Air Force Cambridge Research Laboratories. The stated purpose was "To investigate the automatic recognition and processing of speech signals." Research work conducted from January 1962 through December 1964 is reported here. Mr. Philip Lieberman, CRBV, Hanscom Field, Bedford, Massachusetts, was contract monitor.

Dr. George W. Hughes, Associate Professor of Electrical Engineering, Purdue University, was the Principal Investigator. Much of the material contained in the report was written by Dr. John Hemdal[*] and appeared in his doctoral thesis, "The applications of distinctive features to the primary recognition of speech," which was submitted in June 1964. Some of the work of graduate students Thomas B. Snow (low-frequency harmonic analysis) and William G. Ely (stop consonants) is included in this report.

Other partial presentations of this research have been made before the Acoustical Society of America (see appendices I and II), and in the Proceedings of the Symposium on Models for the Perception of Speech and Visual Form (in press).

[*] Presently: Research Associate, Center for Research on Language and Language Behavior, University of Michigan.

ABSTRACT

The limitations of speech recognition procedures which depend solely on acoustic data are discussed. One such "primary recognition" scheme, based on phoneme classification by tracking the acoustic correlates of a set of distinctive features, is presented. Programmed on a digital computer, these logical operations on digitalized spectra of 17-msec samples of speech were tested on some 300 nonsense utterances from two talkers. A priori information about individual talker characteristics is incorporated into the logic (single-speaker approach). Comparison of machine performance was made with both the intent of the speaker and with the judgments of listeners. Listeners were presented with the same acoustic stimuli that were machine processed. Some perceptual tests were run on short vowel segments excised from nonsense syllables.

Detailed quantitative results are presented only for vowels. They show that man and machine agree about 90% of the time on vowel judgments under these conditions of minimal contextual information. Clear feature boundaries are shown on the F1-F2 plane for the (stressed) vowel utterances. Although these boundaries are not generally valid for more than one voice, simple translations of them may suffice to obtain usable vowel separation for many talkers.

TABLE OF CONTENTS

LIS OF ILLUSTRATIONS

Chapter 1

INTRODUCTION

A. Objective

Our principal objective was to perform research in speech analysis with particular regard to automatic speech recognition. This work follows the combined linguistic-acoustical phonetic approach begun by Jakobson, Fant and Halle (14) and continued by Halle (7, 3), Fant (4), Hughes and Halle (13), and Hughes (12). Certain limitations, however, were placed upon this study in order to obtain more explicit results at the expense of generality. For example, both a restricted context and an analysis of each individual speaker's embodiment of distinctive features (here called the single-speaker approach) were imposed.

B. General Background

We study the physical properties of the speech waveform, which give rise to the distinctive features of an utterance, and the physiological correlates which characterize these properties. A given articulation results in a unique acoustic waveform; but the transformation is not one-to-one, in that one acoustic property may arise from more than one articulatory configuration. The speech signal contains certain properties assumed to be measureable, which serve as the invariants of speech; a change of any one of these properties produces a different utterance. The phonemic description is a complete one; every possible speech utterance in a given language finds expression in terms of the phonemic entities and their phonetic symbols.

Apropos of the general speech recognition problem, two facts have emerged clearly in the past five years: (i) Processing the acoustic signal will not yield a phonemic transcription of an utterance without analysis including at least stress patterns and, most probably, syntax (20). (ii) Any effective analysis scheme will include structural feedback, that is to say, will be based on "analysis by synthesis" 25 . It is, therefore, important that the scope or intent of any "speech recognition" scheme be delineated clearly at the outset.

It is profitable to divide language perception into orders of complexity or into levels of recognition. These may range from those involving concepts and meaning, through those including knowledge of morpho-phonemics and syntax, on down to reception and tracking of quasi-independent acoustic features. This last or lowest level we choose to call "primary" recognition. It alone forms the subject matter of this report.

Although any recognition scheme that aims at completeness almost certainly will incorporate analysis-by-synthesis, we feel that at the lowest or primary recognition level it is both unnecessary and overly expensive. This conclusion is based primarily on our observation that further improvement in acoustic feature teaching will not materially improve machine recognition "scores", nor will it increase the number of distinctive features detectable. Analysis-by-synthesis techniques have made possible the accurate tracking of pole-zero locations 24, 1) which is important to the problems of articulation or production. It appears, however, that simple "open-ended" analysis of filter outputs (peak picking, for example) does extract sufficient information about vowel and sonorant formant frequencies, intensity temporal variations, presence of voicing, presence of turbulence, and even fundamental pitch to allow programmed recognition procedures to "score" as well as panels of listeners presented with identical isolated speech segments.

Since speech is considered by some to be an "overlaid function" in the human organism, it is to be expected that many properties of the speech waveform would not be of value in determining "what was said." That is, the acoustic waveform is produced by modulating an expulsion of air by the movement and placement of certain organs which are primarily used for the intake of food and oxygen. Furthermore, this modulation and coding by means of the language is not done most efficiently, and there are many superfluous properties of the resulting acoustic signal which do not serve to indicate the content of the message. Although the speech signal contains a large degree of redundancy and hence resistance to noise, some of these waveform characteristics are completely extraneous, and others allow the listener to perceive information other than the "message." It is necessary to carefully distinguish between those properties common to the "linguistic production" for all speakers of the language and those properties of speech which may be peculiar to each individual, viz., the physiognomic features, the configurational features (2) and the singular features,* reviewed below:

The physiognomic or expressive features of speech signal the emotional attitude of the speaker and therefore may be distinct from the literal content of the utterance. These features may convey to the listener that the speaker is angry, happy, sad or in some other emotional state. For example, the intended meaning of the two questions, "Why don't you go jump in the river?" (in order to cool off) or "Why don't you go jump in the river?" (because you're bothersome), may be made clear to the listener by a change of the physiognomic features of the two utterances.

The configurational features of the speech utterance are those properties which indicate division into the various grammatical units of the language. These features would enable the listener to determine whether "I scream" or "ice cream" had been intended. For example, these features might be exhibited as changes in stress, pitch, tempo, etc.

The singular features are those properties residing in the speech waveform which inform the listener of such possibilities as the identity of the speaker, his location, his physical condition, sex, age, health and perhaps other qualities of an indicative nature. After a little practice, it is possible to recognize a person by the sound of his voice, and one can usually distinguish male and female speakers

* Referred to as "idiosyncratic" features by Ladefoged and Broadbent, Reference 16.

on first hearing a small sample of their voices. Speaking into a rain barrel or over a telephone changes the "quality" of the voice in such a way that these situations can be readily recognized. All of these differences may be attributed to changes in the singular features contained in the acoustic signal.

Although each of these features is acquired by the speaker, nevertheless they are distinct from the content of speech. This concept of the content of speech, in particular the information-bearing aspects of the speech waveform, needs to be carefully delineated from the physiognomic, configurational and singular features of speech.

In receiving and recognizing speech, the listener extracts from the speech waveform a certain amount of information from which he determines "what was said." Combined with this extracted infoımation are many other informational cues which come from a reservoir of semantic and linguistic additives in the listeners' memory and which are stored in the process of learning the particular language. This combination enables the listener to approximate the same mental images and forms as the talker. It should be emphasized that the phonic data will not provide all the clues used by the listener in identifying an utterance. The interpretive faculty of the listener comes into play to a large extent in normal speech, combining knowledge of the speaker and the language. Thus, rules of syntax, grammar and morphology will provide clues for correcting perception errors at levels other than the recognition of the sound shapes. There is reason to believe that as more familiarity of the speaker and his subject is gained by the listener, less use is made of the acoustic waveform 22.

We have adopted the viewpoint that it is more interesting to see how much information can be extracted from the speech waveform than it is to study how few phonic data are necessary in normal conversation. What is the maximum information inherent in the physical signal, and how is it coded in the waveform? The recognition of unfamiliar proper names is an example of the necessity of deriving all the information from the physical waveform, since context and learning are not involved, at least at the word level. The correct recognition of nonsense words also must be based entirely on the acoustic signal. It is for this reason that the initial computer recognition studies undertaken for these experiments dealt with simple nonsense forms. In this way, the computer results could be compared with results of listening tests on the same speech data. Thus, neither machine nor human listener was able to make use of rules of syntax, grammar and morphology, or context and meaning. The comparison, then, is a fair one.

Consider the following thought experiment. A sequence of nonsense utterances consisting of possible English forms but lacking any grammatical construction is presented to a listener who correctly copies their written equivalents. A second speaker then reads the written forms, and a second listener attempts their recognition. The common core throughout this process is the "content" of the original utterances, and the recognition of this common core of information is termed the _primary_ recognition (5). The physiognomic, configurational and singular features are not common to all levels throughout the various transformations. The purpose of mechanical speech recognition is to automatically perform at least the primary recognition.

This description of primary recognition has as its purpose a careful determination of the limits and goals of mechanical speech recognition. For example, since speech contains a high degree of redundancy, the next sound following a known sequence of sounds may in many cases be partially or wholly inferred. The sequence /ʃʌ/ has a high probability of being followed by an /n/ and therefore this redundancy may be used to reduce the quantity and quality of the physical measurements used to detect the /n/. There is no doubt that any recognition scheme would benefit from the application of the knowledge of the frequency of occurrence of the sounds of normal speech. However, in view of the definition of primary recognition, the use of such digram or trigram probabilities is not a proper part of primary recognition, since this recognition must take place at a level which includes arbitrarily-determined nonsense forms where these probabilities are no longer applicable.

This limitation on the uses of information concerning the frequency of occurrence of the sounds in a given language is not to be confused with certain linguistic constraints imposed by the particular language, and which may properly be included in the problem of primary recognition. Although nonsense utterances were used in the thought experiment previously described, they must comply with the allowable forms of English. Thus, there will be no initial /ʒ/ or final /h/ in the list of nonsense forms, nor will there be any other sounds or sequences of sounds in positions or combinations which are not found in normal English. For example, there occur no consonantal clusters in English which have both tense and lax phonemes. This constraint should be applied at the level of primary recognition even though a statement of this constraint would be identical to a possible probability statement, such as "/s/ follows /d/ with zero probability if there is no intervening word boundary."

YET	/jεt/
HANG	/hæŋ/
POD	/pad/
CHUM	/tʃʌm/
BOUGHT	/bɔt/
LIEGE	/liʒ/
RIG	/rIg/
FIRTH	/fɚθ/
COOK	/kʊk/
SOOTHE	/suð/
WAVE	/wεIv/
SHOWS	/ʃoʊz/
I	/aI/
JOIN	/dʒoIn/

Figure I-1. A table of English words and their transcriptions into phonetic symbols. This list illustrates approximately what speech sounds (in midwestern dialect) correspond to the notation used in this report.

Chapter 2

GENERAL PROCEDURES

A. The Single-Speaker Approach

It has been pointed out by Fant (4) that the physical properties of each distinctive feature may be relative for each speaker, and that proper normalization must be performed on the speech data in order to adequately track the features. A useful method of normalization has not been determined. It has been suggested by W. Lawrence (17) that "...all formant frequencies are judged in relation to the long term mean value of the formants for that speaker. When listening to the first few words of a speaker we quickly estimate these mean values and thereafter judge all sounds relative to them." It is reasonable to expect that a mechanical device would be required to perform automatically a similar sort of normalization in order to handle the diversities of several speakers. In view of this, the experimental approach used here has been designed to lay the groundwork for a study of the necessary normalization. It is proposed that by adjusting a recognition scheme to the speech of a single individual and then introducing the speech of other persons into the same program, the minimum changes necessary to adapt the program for the other persons may be determined. These minimum changes are related to the prior information that the recognition device needs in order to achieve the adaptation The possibility of having the computer automatically perform the necessary changes on the basis of a pilot utterance should not be overlooked.

The single-speaker approach has two possible advantages. First, rather than performing more complex and broad measurements in an effort to contend with the vagaries of multi-speaker utterances, the tests and logic of the recognition device are adjusted in an optimal manner to the speech of one individual. If, then, the speech of another individual is introduced into the same system, the distribution and types of additional errors incurred with the new speech are directly related to the necessary and sufficient information from which changes in logic, measurements or thresholds may be determined in order to adjust optimally the program to the new speaker. It is obvious that only those measurements producing errors need to be considered for possible change, and this approach determines just those error-producing measurements. In addition, the changes required may, in many cases, be correlated with physiological differences in the vocal apparatus of the speakers or with certain differences in articulation. For example, Fant (3) points out that females average 10 to 15% shorter vocal tract lengths and a corresponding increase in the difference between the first two vowel formants. The incorporation of this information in adjusting the measurements and thresholds of a program designed for a male voice would improve the recognition ability of the device when confronted with the speech of a woman.

The second advantage of this approach results from an examination of the necessary changes. If the rule of simplicity is imposed on the distinctive features, then the rule of simplicity must also operate on the adaptation changes which are required in the program. Simple shifts in various thresholds would imply a "correctness" in the tests and measurements, while a complete change in certain procedure would imply that the proper physical correlate of the distinctive feature had not been determined. To illustrate, suppose speaker A has a pronounced energy peak at about 500 cps in every vowel which immediately precedes a nasal consonant. The detection of this peak provides a highly reliable indication of the feature nasality in the following consonant. However, this peak is only concomitant to nasality in speaker A. A program which performs this measurement to indicate nasality in consonants would fail with speaker B whose vowels lack this 500 cps peak when preceding nasal consonants. Furthermore, a simple change in threshold, say from 500 to 600 cps, would not suffice to improve the results with the speech of speaker B; a completely different physical parameter would have to be tracked in order to adjust the program to the new utterances containing nasal consonants. The implication of this result is that the 500 cps peak of speaker A is not the physical correlate of the feature nasal/oral. If, on the other hand, a simple change yields a large return in the accuracy of the results when new speaker data are introduced, the implication serves to strengthen the supposition that the parameter being tracked is a proper correlate of the distinctive feature.

Finally, since virtually all our recognition or feature-tracking procedures were programmed for a high-speed, general-purpose digital computer, it is appropriate to specify the role of this machine in our research. In this study, no learning or adaptive procedures were programmed. No attempt was made to obtain statistical threshold results based on large amounts of data. The computer, as we used it, served as a convenient tool for modeling feature-tracking schemes based on spectral data manipulation and deterministic decisions.

B. A Feature Framework

The description of speech proposed by Jakobson et al. is in terms of a given set of binary distinctive features, their physical properties and their articulatory correlates (14, 15). These dichotomous features are a minimum set of purely relative properties of speech that the listener needs in order to distinguish among all except homonymic morphemes - whenever the distinction is possible - without help from context or situation. This particular set of distinctive features was proposed because it afforded a simple description of speech at several linguistic levels including the perception of speech.

Because of the desire to achieve great generality in the application of the distinctive feature approach to the process of speech, the ease of measurement of the physical properties of the distinctive features was not a consideration in this description. In fact, "Theoretical constructs are never introduced because of considerations that have to do with analytic procedure" (6). Therefore, an engineering approach to the tracking of the distinctive features has suggested several minor changes in Jakobson's original formulation in order to take advantage of simpler measurement procedures. For example, /k/ and /tʃ/ are both compact as opposed to diffuse and one physical correlate of compactness may be its higher energy (average power times the duration). However, the energy of the /tʃ/ is usually two or three times that of a /k/ while the range of energy of the /k/ and the /t/ often overlaps (although on the average a /k/ is "stronger" than the /t/). Thus, if relative energy is used as the measure of compactness, it would be convenient to consider only the /tʃ/ as a compact, tense and interrupted phoneme. Other expediencies will be described later, along with the details of each of the physical measurements.

Figure II-1 describes 34 phonemes in terms of their distinctive features. Various authors put the number of phonemes of English at 28 to 50 with no complete agreement. This disagreement results from different ways of handling variants of phonemes and some complex two-valued speech sounds (which may or may not be transcribed as a single phoneme), the level of speech considered, and the purpose for which the description is to be used. (Phoneticians usually list the largest number of phonemes probably in an effort to make precise the slight differences in pronunciation due to various talker dialects.) The automatic recognition of speech by machines also imposes a criterion for evaluating the "correctness" of a phonemic representation; e.g., the sound /e/ as in gate and weight may better be transcribed as /ɛI/, a diphthong, since this form is more closely associated with the physical measurements. In other words the machine, when confronted with this sound, requires a simpler rule to produce the /ɛI/ transcription than /e/.

Each column in Figure II-1 contains a sequence of pluses and minuses and blanks which are a unique specification for the phoneme indicated. The plus sign indicates the presence of the first named feature, a minus sign indicates

DISTINCTIVE FEATURES	PHONEMES																																	
	ɚ	i	I	ɛ	æ	a	ʌ	ɔ	U	u	j	r	w	ℓ	m	n	ŋ	ʃ	s	f	θ	ʒ	z	v	ð	tʃ	k	p	t	dʒ	g	b	d	h
1. Sonorant / Nonsonorant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
2. Consonant/Nonconsonant	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−
3. Continuant/Interrupted																		+	+	+	+	+	+	+	+	−	−	−	−	−	−	−	−	
4. Nasal/Oral											−	−	−	−	+	+	+																	
5. Tense/Lax	+	+	−	−	+	+	−	+	−	+								+	+	+	+	−	−	−	−	+	+	+	+	−	−	−	−	
6. Compact/Diffuse	−	−	−	+	+	+	+	+	−	−					−	−	+	+	−	−	−	+	−	−	−	+	+	−	−	+	+	−	−	
7. Grave/Acute	−	−	−	−	−	+	+	+	+	+	−	−	+	+	+	−		−	−	+	+	−	−	+	+			−	−			+	−	
8. Flat/Plain	+	−				−		+			−	+	+	−																				
9. Strident/Mellow																				+	−			+	−	−	−			+	−			

Figure II-1. Distinctive Features of the Phonemes of English.

either its absence (or the presence of the complementary feature), and neither sign indicates that this feature is irrelevant to the uniqueness of the representation. The tree diagram of Figure II-2 portrays the binary nature of each feature chosen and the order in which the measurements of each feature is performed.

This diagram differs in slight detail from those of other sources. For example, the tree diagram in Figure II-2 is different from that of Jakobson, Fant and Halle (14). The differences and changes between these two diagrams were made in order to present the features of English in a simpler and more unified manner. It should be noted that, with two exceptions, Figure II-2 shows the same feature to be operating on each level, regardless of the preceding level. Thus, at the fifth level, the feature grave/acute is applicable for both vowels and consonants.

The decision tree and the order in which the features are applied were designed to take advantage of simpler measurement procedures. It is not being suggested that those changes which facilitate the physical measurements may also afford simplification on other levels of linguistic analysis. However, one change (/e/ - /ɛI/ and /o/ - /oʊ/) may allow reduction of the number of distinctive features operating on the vowels by one. The compact/non-compact and diffuse/non-diffuse distinctions are reducible to a single feature, compact/diffuse. In order to see how a reduction in the number of features is obtained, it is necessary to give an operational distinction between two kinds of vowel-like sounds, the pure vowels and the diphthongs.[*] A pure vowel is defined as a sonorant, non-consonantal sound, which is capable of being sustained indefinitely, while a diphthong would not be recognized properly if it were sustained.[**] A sound spectrogram would show constant formant frequencies for the pure vowels (excluding formant

[*] In many respects the description which follows parallels that of K. Wiik (26). However, the purpose for the description is different, and thus some details are different.

[**] This definition differs from that of Lehiste and Peterson (18) who consider three vowel-types: the pure vowel, the complex vowel and the diphthong.

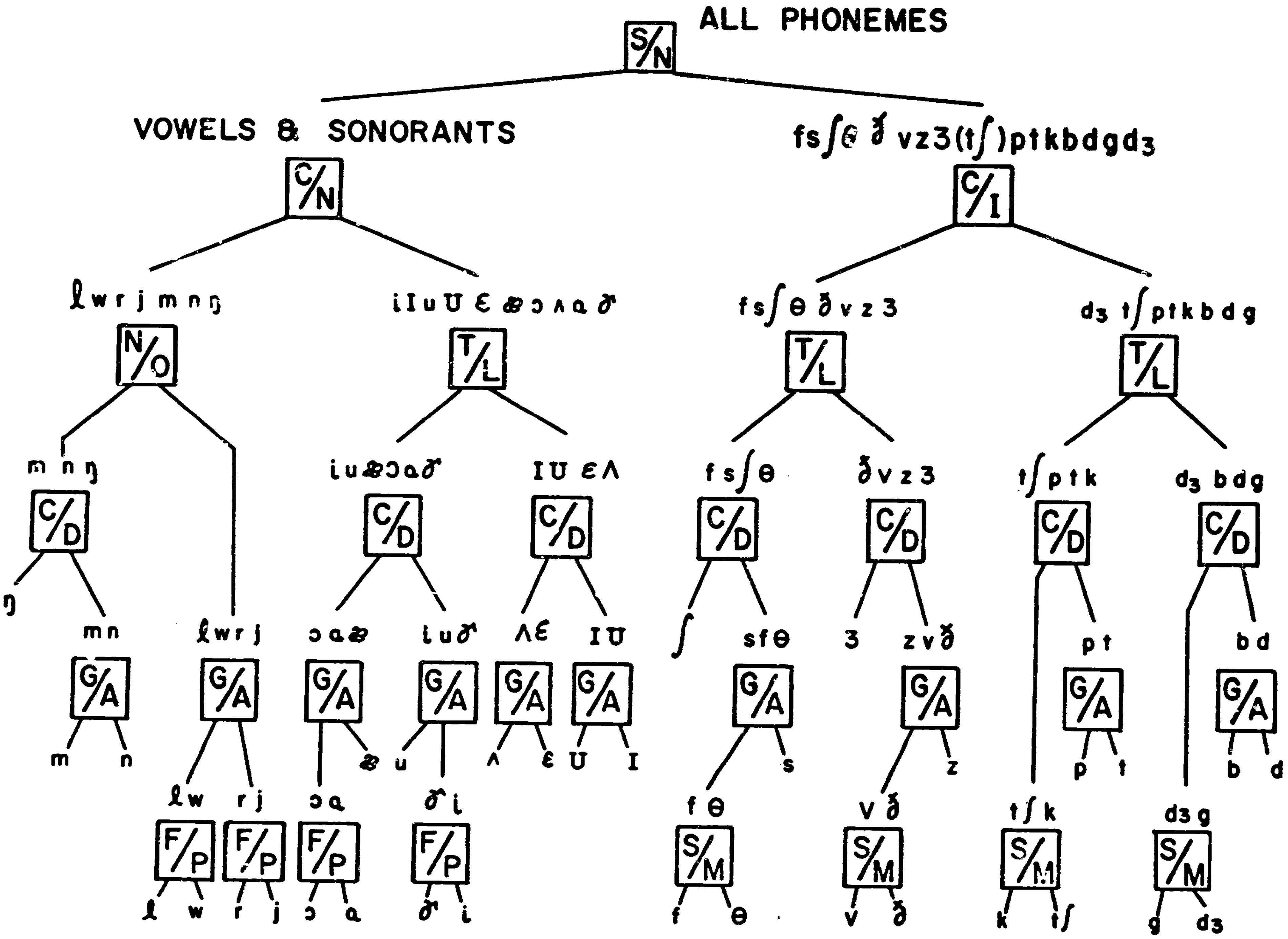

Figure II-2. Decision Tree.

transitions due to the consonantal environment), while the variable vowels or diphthongs would exhibit a changing frequency for one or more formants. Now, if we assume for the moment that all pure vowels may be distinguished on the basis of only the frequencies of the first two formants, then these vowels are characterized by a single point in the F1-F2 plane (a plot of the first formant frequency against the second formant frequency). Of course, isolated vowels which are judged the same by a panel of listeners would not result in identical points in the F1-F2 plane but would form a cluster of points in a relatively small region of the plane. It follows, from our assumption, that the first two formant frequencies are sufficient to characterize each vowel and that there will be no overlap of regions representing two different vowels - only a continuous change in the amount of agreement among the listeners as to the identity of the vowel. A boundary between vowel types is defined as the locus of points representing the formant frequencies of vowels which are judged by half the listeners as one vowel and by the other half as an adjacent vowel. Vowels are said to be adjacent on the F1-F2 plane if the change in agreement among the listeners changes uniformly from one vowel to another as the formant frequencies change along a path perpendicular to a boundary. If the path of formant frequencies leads from a region of uniform agreement to a region where no agreement is possible, than the latter region is called a hole in the F1-F2 plane; and it represents formant frequencies which are not found in English vowels. It is clear that if the listeners are operating under forced choice conditions, then the entire F1-F2 plane is covered by regions representing vowel types or holes. The formation of the vowel-vowel boundaries or the vowel-hole boundaries is more easily visualized if one considers the speech sounds to be synthesized vowels of two formants, since these formant frequencies may be continuously shifted. The boundaries need not be the same as those determined with natural vowels, but the principle is the same. If, when crossing a boundary from one vowel to another, the agreement in the classification does not proceed uniformly from the first vowel to the second (thus implying vowel region overlap), then the assumption that only the first two formants are sufficient for recognition is not valid. This occurs if the vowel /ɚ/ is included, since the lower third formant is an important cue for this vowel. However, if we consider regions and boundaries in an F1-F2-F3 space, then no overlap will occur. We should also add that any additional parameter necessary to distinguish vowel types, such as the fundamental frequency of the voice excitation, or duration, merely requires another coordinate in the parameter space.

Given the description made thus far, we would expect that although small formant variations may occur in the pronunciation of constant vowels, nevertheless these variations trace out paths which are wholly contained within the given vowel boundaries. The path of the formant variations of diphthongs, on the other hand, would not be expected to lie within a single vowel boundary but would cross one or more boundaries. Furthermore, it is just this change in classification as the formant path crosses the boundary that completely characterizes diphthongs.

Spectrograms of words containing the vowel sounds in the words <u>weight</u> and <u>home</u> frequently indicate that the second formant in both of these sounds is continually changing frequency throughout the entire length of the sound. Consequently these sounc would trace out paths in the F1-F2 plane which would cut across other vowel boundaries. Therefore, we will consider these two sounds as diphthongs and transcribe them as /ɛI/ and /oʊ/, respectively. There are three advantages to considering these two sounds as diphthongs: 1) the logic requires a simpler test for producing /ɛI/ anc /oʊ/ transcription than /e/ or /o/; 2) it allows a reduction in the number of distinctive features necessary tc describe the vowels; 3) if this change is not made, it would be necessary to allocate the proper region in the F1-F2 plane for these vowels. However, no "holes" exist in the F1-F2 plane where these vowels would be expected to appear.

That is, neither F1 nor F2 nor duration serves to separate these vowels from all others, and indeed it seems to be the change of the second formant which performs this task.

Each vowel, which is described by the given set of distinctive features, must be consistent with the physical correlates which are known to be applicable to these features. The vowel /ɚ/, for example, is now described as acute, diffuse, tense and flat. Among other measures of acuteness, Jakobson suggests, ceteris paribus, that the second formant lies relatively closer to the third than the first; thus the ratio $\frac{F2-F1}{F3-F2}$ should be larger for acute vowels than the corresponding grave ones. Because an /ɚ/ has generally a lower third formant than the other vowels, this ratio tends to be larger. A comparison of this one measure of acuteness shows that the /ɚ/ may be considered acute:

/i/	2.80	/u/	.416
/I/	2.86	/ʊ/	.475
/ɛ/	2.05	/a/	.267
/æ/	1.54	/ɔ/	.172
/ɚ/	2.53	/ʌ/	.458

This comparison is based on average formant frequencies given by Peterson and Barney (23).

Furthermore, the /ɚ/ is diffuse. This feature is primarily characterized by a low first formant. The location of the /ɚ/ on an F1-F2 plot in Figure II-3 shows that it lies to the left of the tentatively selected diffuse-compact boundary, along with all other diffuse vowels /u/, /ʊ/, /i/ and /I/.

The /ɚ/ also exhibits the proper tenseness by virtue of its greater energy and duration.* It is our view that the question as to whether /ɚ/ should be

* House 11, has suggested that if /ɚ/ be taken as lax the symmetry of his vowel duration relation improves. This, however, seems to be true only with ordering of the vowels as given by House. In his diagram he has placed and thus seemingly compared the /ɚ/ duration among the vowels /æ/ and /a/. Therefore, since /ɚ/ is relatively short compared to /æ/ and /a/ and hence lax,

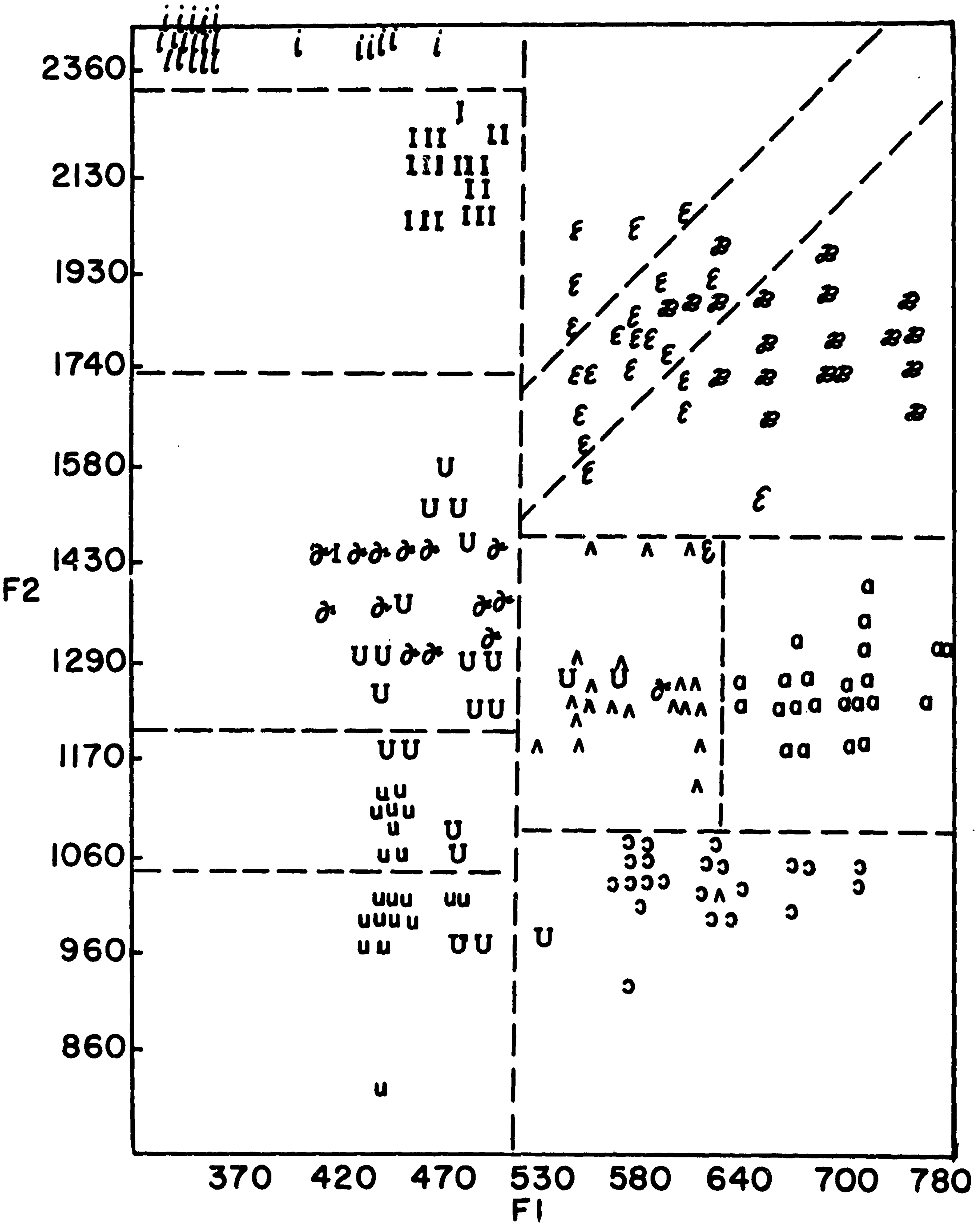

Figure II-3. Formant Locations Speaker GWH.

considered tense or lax is best decided on the basis of comparing the physical measure of the tension between /ɚ/ and the phoneme with which it is in opposition. Examination of the idealized F1-F2 plot of Figure II-4 shows that there is probably no phoneme which differs from /ɚ/ in the tense/lax feature alone. Therefore, we seek that feature which is a secondary influence on vowel duration. Evidence indicates that the diffuse (close) vowels are shorter, ceteris paribus, than the compact vowels. We should then compare the duration of /ɚ/ with the other diffuse vowels /i/ - /I/, /ʊ/ - /u/. On the basis of this ordering, we find the duration of /ɚ/ prescribes a tense vowel. If symmetry is desired, this manner of comparison provides it.

Again, examination of Figure II-4 shows ten vowels. Three features (grave/acute, flat/plain, and compact/diffuse) are based on formant location alone and divide the ten vowels into equal groups of five each. However, there are, according to this scheme, only four lax vowels and six tense vowels. This does not upset the symmetry, however, since only vowel pairs which differ in the feature tense/lax (i.e., /i/ - /I/, /æ/ - /ɛ/, /ɔ/ - /ʌ/ and /u/ - /ʊ/)also exhibit the other physical manifestation of the tense/lax tendency - namely, a shift toward a neutral or ideally lax formant location. This tendency is shown in Figure II-4 by the near intersection of the lines drawn from the tense vowel region through the region of its lax vowel counterpart.

According to this scheme, the /ɚ/ must be flat rather than sharp. Jakobson suggests that sharpness consists of a tendency of all formants to be higher relative to the flat vowels. It, of course, is necessary to compare the /ɚ/ with the /i/, since only these two must be separated at the level at which the sharp/flat feature operates. Although the first formants of the two vowels are nearly the

House's ordering of the vowels yields the sequence:

tense	lax	tense	lax	tense	lax	tense	lax	tense	tense	lax	tense.
/i/	/I/	/e/	/ɛ/	/æ/	/ɚ/	/a/	/ʌ/	/ɔ/	/o/	/ʊ/	/u/

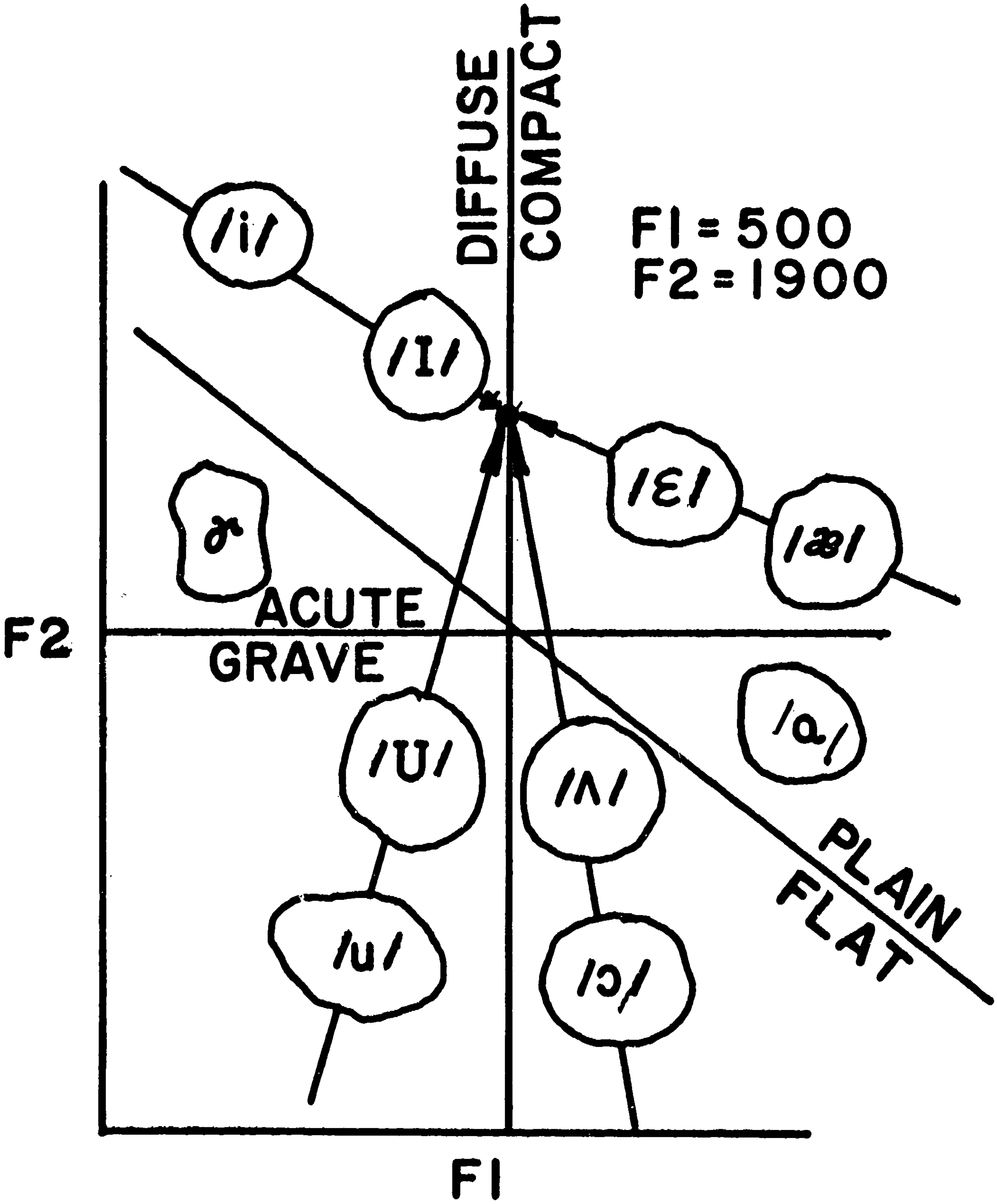

Figure II-4. Idealized Formant Locations for Vowels.

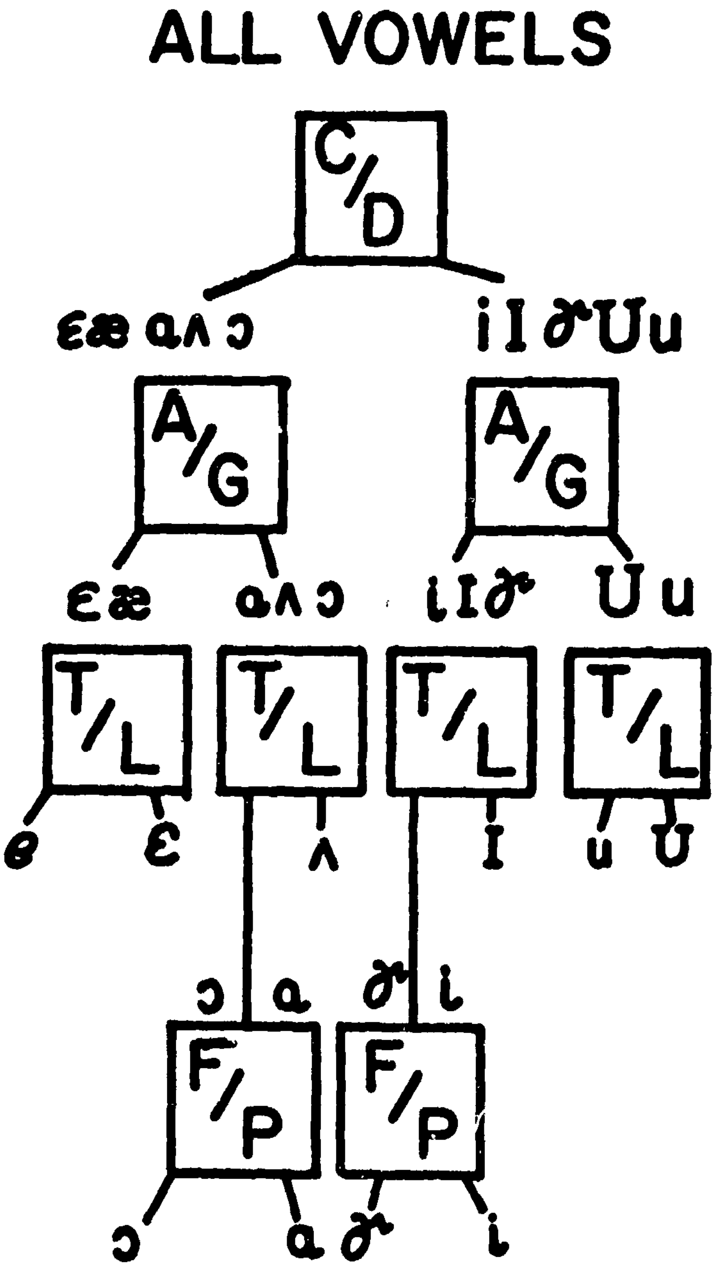

Figure II-5. Decision Tree for Vowels.

same frequency, all higher formants of the /ɚ/ are lower than the /i/, thus qualifying the /ɚ/ as a flat vowel.

This reduced set of distinctive features applying to the vowels has as its main advantage a simplicity both on the conceptual level and the level of the physical measurements. It will not necessarily produce a simplicity at other levels of linguistics. However, the four advantages listed below make this particular decision tree attractive for the purposes of machine recognition.

First, the decision tree exhibits a symmetry which allows a minimum of distinctive features to describe each vowel. (The sounds /o/ and /e/ have been excluded from the tree and are to be treated as suggested earlier as diphtongs.)

Second, there is a distinct possibility that the physical correlates of each feature in this set are independent; i.e., the order in which the measurements are performed on the vowels will not alter the form of the measurements nor the various thresholds. This property is illustrated in Figure II-4. An idealized F1-F2 plot shows the average locations of the ten vowels. Three curves represent, respectively, the boundaries between acute/grave, compact/diffuse, and flat/sharp. If the tense/lax feature is primarily measured by duration, then a third axis out of the paper labeled "length" will allow the tense vowels to be separated from the lax vowels by a plane lying somewhere above the plane of the paper. This boundary plane need not be parallel to the plane of the page; it will, in fact, slope toward the close (diffuse) vowels as indicated by House. (See Figure II-6). These same boundaries will separate each vowel regardless of the order in which the features are detected. The decision tree of Figure II-5 shows one possible permutation of features.

Third, the intersection of all boundaries may indicate both conceptually and actually the location of the ideal neutral vowel.

The fourth advantage stems from the fact that only four simple boundaries need to be changed when adapting a vowel recognizer to the speech of a different individual. One additional note: Each of the features (three), which are shown as boundaries on the Fl-F2 plane, separates the vowels into groups of five each. Also, without the added dimension of vowel duration, there is some overlap found among vowel groups. With the added duration axis this overlap disappears almost entirely.

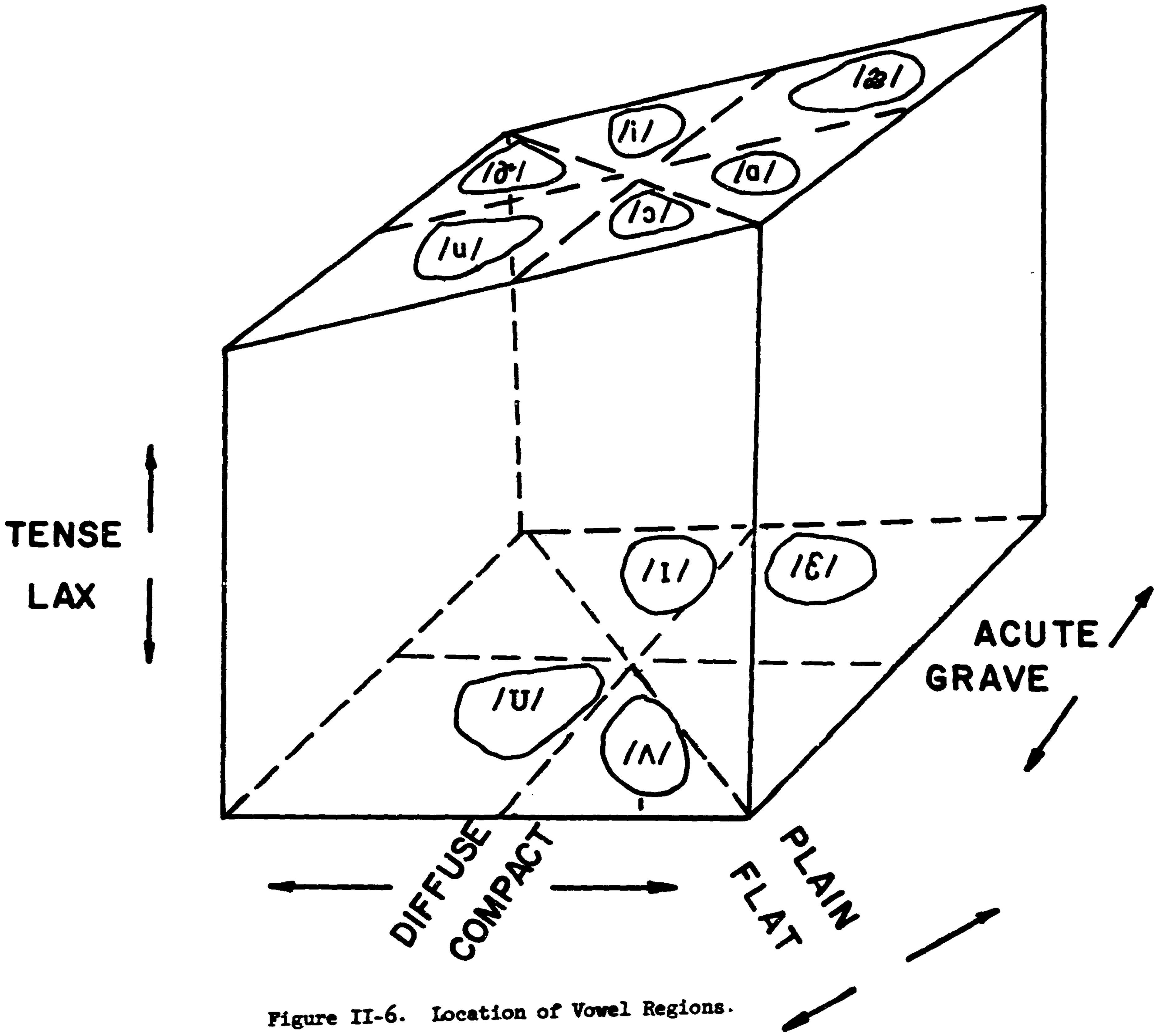

Figure II-6. Location of Vowel Regions.

Chapter 3

PROGRAMMED RECOGNITION PROCEDURES

A. General Segmentation

Operations performed on the speech data by the computer fall into two groups: those which may be performed by examining properties found in a single sample or time segment, and those which require the examination of the combined characteristics of several segments or which depend upon previous decisions performed on several surrounding time segments. The former group is performed initially on each segment in sequence, while the latter often requires several passes on the data.

The first decision sequentially classifies each time segment as belonging to the speech utterance or to a silence by detecting a significant amount of energy above the noise level of the system. Each time segment not classified as silence is then labeled as belonging to either a sonorant or a nonsonorant phoneme. This is done, as shown in the block diagram of Figure III-1, by first eliminating segments belonging to stop, fricative and affricate classes of phonemes. Thus, if any time segment has one or more of the following properties, it will be classified as belong to a nonsonorant phoneme:

1) A lack of voicing (as measured by the absence of energy at frequencies below 350 cps.)
2) The presence of indications of turbulence (as measured by the presence of frequency components above 4000 cps.)
3) The presence of silence, with or without voicing, which precedes the segment under consideration.

Separating the nonsonorant time samples in this manner results in catagorizing the stops, fricatives, and affricates into four separate groups as shown in the four blocks in the upper right of Figure III-1. Thus, property 1) above may be considered as the physical correlate of the feature tense-lax[*], and properties 2) and 3) are to be associated with the feature continuant/interrupted.

The time segments labeled sonorant are further classified as belonging to either consonantal or nonconsonantal phonemes. This classification, dividing the vowel segments from the non-vowel sonorants, is based on the prope~ty that the non-vowel sonorants are weaker sounds than the neighboring vowels, due to a more complete closure of the vocal tract. A downward change in spectral energy with time is characteristic of a vowel/non-vowel boundary, while an increase in energy signals a non-vowel/vowel boundary.

B. Vowels

1. Formant tracking

It has been recognized that most of the information conveyed by vowels is contained in the location of the first two or three formants (poles of the transfer function of the vocal tract). Therefore, after determining that a given time segment belongs to a vowel, a computer routine is entered which attempts to locate the frequency of the first three formants by locating the spectral peaks in each of the first three formant regions. There exist certain irregularities in the location of these peaks with time as a result of momentary breaks in voicing, small amounts of extraneous noise, and other sources of nonmeaningful variations. A very simple constraint taking the form of peak-smoothing is all that s necessary to compensate for a majority of these fluctuations. Momentary dropouts are eliminated by requiring that each formant positior be equal to or lie between the positions determined for the preceding and following segments. Long term dropouts, unaffected by this simple smoothing, may be handled more efficiently at higher levels of the

[*] For the English nonsonorants, voicing is only concomitant to the tense/lax feature, but it is nevertheless usually reliable and easily measured and, therefore, it will be used as if it were the correlate itself. Its application as a measure o laxness may be overridden as additional information warrants.

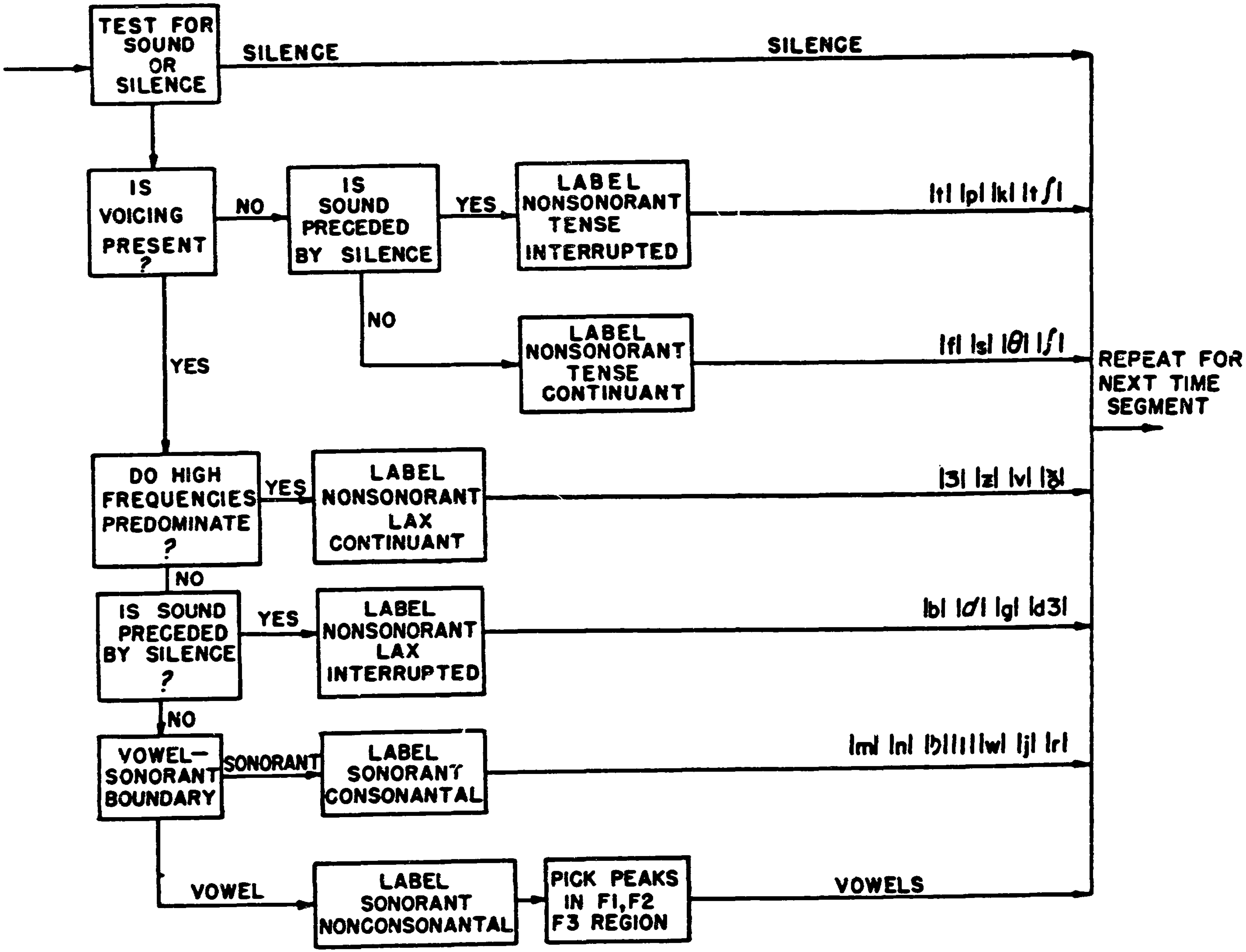
TEST FOR SOUND OR SILENCE
SILENCE
SILENCE
IS VOICING PRESENT ?
NO
IS SOUND PRECEDED BY SILENCE
YES
LABEL NONSONORANT TENSE INTERRUPTED
|t| |p| |k| |tʃ|
NO
LABEL NONSONORANT TENSE CONTINUANT
|f| |s| |θ| |ʃ|
REPEAT FOR NEXT TIME SEGMENT
YES
DO HIGH FREQUENCIES PREDOMINATE ?
YES
LABEL NONSONORANT LAX CONTINUANT
|ʒ| |z| |v| |ð|
NO
IS SOUND PRECEDED BY SILENCE ?
YES
LABEL NONSONORANT LAX INTERRUPTED
|b| |d| |g| |dʒ|
NO
VOWEL— SONORANT BOUNDARY
SONORANT
LABEL SONORANT CONSONANTAL
|m| |n| |ŋ| |l| |w| |j| |r|
VOWEL
LABEL SONORANT NONCONSONANTAL
PICK PEAKS IN F1,F2 F3 REGION
VOWELS

Figure III-1. Independent Time Segment Classification.

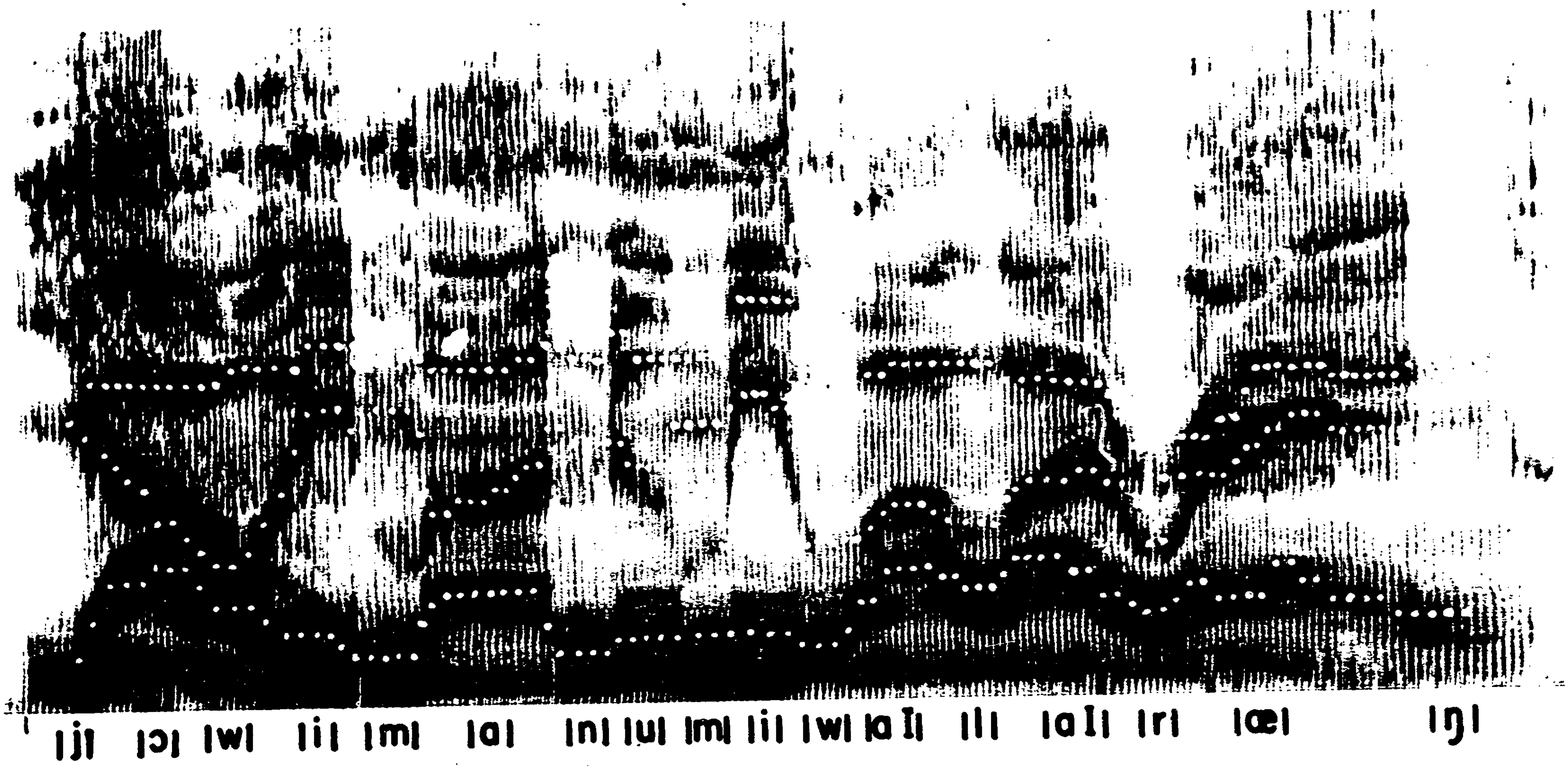

Figure III-2. Example of Results of the Peak Picking Routine.

analysis. A second constraint limits the rate of change of the formant position to a value not exceeding 50,000 cps/second. This maximum allowable change in formant corresponds closely to the maximum change as detected and measured on a sound spectrogram. An example of the results of the peak-picking program is shown in Figure III-2, where the white dots superimposed on a standard sound spectrogram show the location of the computer determined peaks.

2. Feature tracking

Several related measures of the feature acute/grave have been suggested, all of which essentially depend upon the location of the first two formants of the vowels. Jakobson, Fant and Halle (14) suggest a) the center of area, and b) the third moment about the center of area. On the other hand, the relationship

$$\frac{F2 - F1}{F3 - F2} = \text{const.}$$

describes an acute/grave boundary which is more easily programmed and yet provides good separation. The curves of this measure, plotted on the F1-F2 plane, are shown in Figure III-3. Although F3 cannot be shown explicitly on the F1-F2 plane, F3 takes on a nearly constant value of 2400 cps in the region where the boundary is to occur. This particular measure of the grave/acute can be readily programmed into the computer; however, we found it sufficient to consider everything with a second formant greater than 1430 cps as acute and all vowels with a second formant below 1170 cps as grave. For vowels whose second formant lies between these two frequencies, the information concerning the length of the vowel was utilized.

The compact vowels were distinguished from their diffuse counterparts by the location of the first formant. If this formant frequency was above 500 cps (for Speaker GWH), the vowel was classified as compact.

A reasonably good measure of the feature flat/plain is the sum of the first two or three formants. If this sum is high, the vowel is plain; if low, the vowel is flat. However, if this feature is made the last to be measured in a vowel decision tree (relaxing the requirement of independence), it is necessary only to

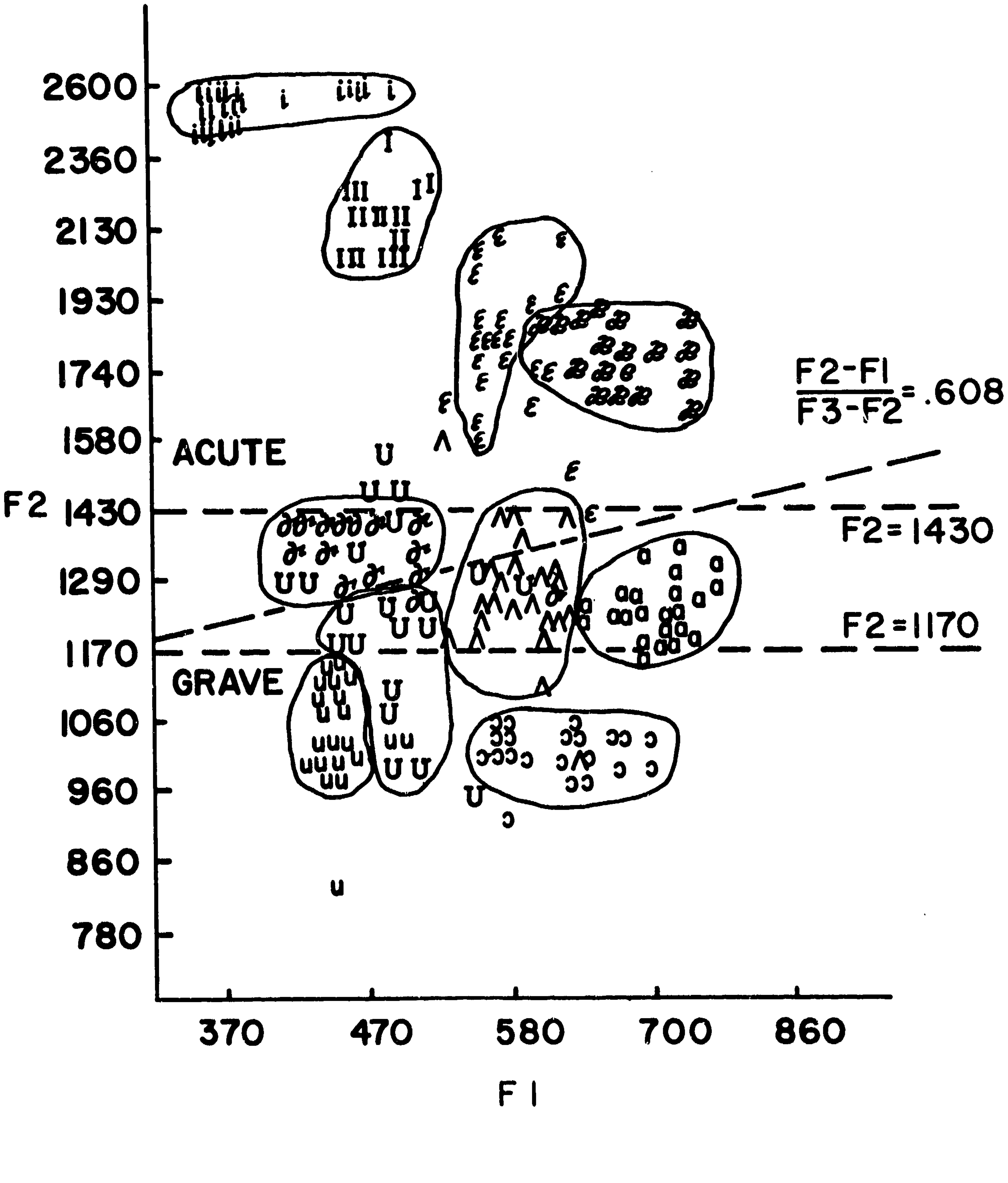

Figure III-3. Grave/Acute Boundaries.

separate /ɚ/ from /i/ and /ɔ/ from /a/. In the first case, /ɚ/ can be more easily separated from /i/ by its lower third formant, and the /ɔ/ can be separated from the /a/ by a lower second formant. This just reflects the fact that a low value for the sum of the first three formants (indicating a flat phoneme) arises predominately from the lower third formant in the /ɚ/ and a lower second in the /ɔ/.

The measure of the feature tense/lax is perhaps the least well known of the vowel features. However, according to Fant (3), one should expect to see a lengthening of the vowel and a shift of the formant positions away from a neutral or most lax position for the tense vowels. In the limited data examined in this experiment, this is precisely what was found, provided that the phonic data were examined for one individual speaker at a time and that the tendency away from neutral was explained on the basis of "true" tense/lax pairs. Figure III-4 shows the distribution of vowel formants in all possible contextual environments (speaker GWH). If lines are drawn from the center of the regions of the four tense vowels through the centers of the regions of the four lax vowels, these lines very nearly converge at an imaginary point somewhere in the neighborhood of F1=500 cps and F2=1900 cps. This convergence occurs only among vowel pairs for which the only distinction is the tense/lax feature. The definition of a "true" tense/lax pair is given as a pair of vowels which differ only in the single tense/lax feature. Fant's description of this feature is valid, then, for "true" tense/lax pairs.

Examination of the speech data of other speakers indicates that this imaginary point of convergence does not occur at the same location for all speakers, nor is it always a well-defined point. Nevertheless, the general tendency remains the same. (See JFH speaker data - Figure III-5.) The data examined thus far seem to bear out the supposition that the physical correlate of tension is a combination of shift of formants away from a neutral position and a relative lengthening of vowel duration.

The tense/lax feature is the only vowel feature whose physical correlate seems

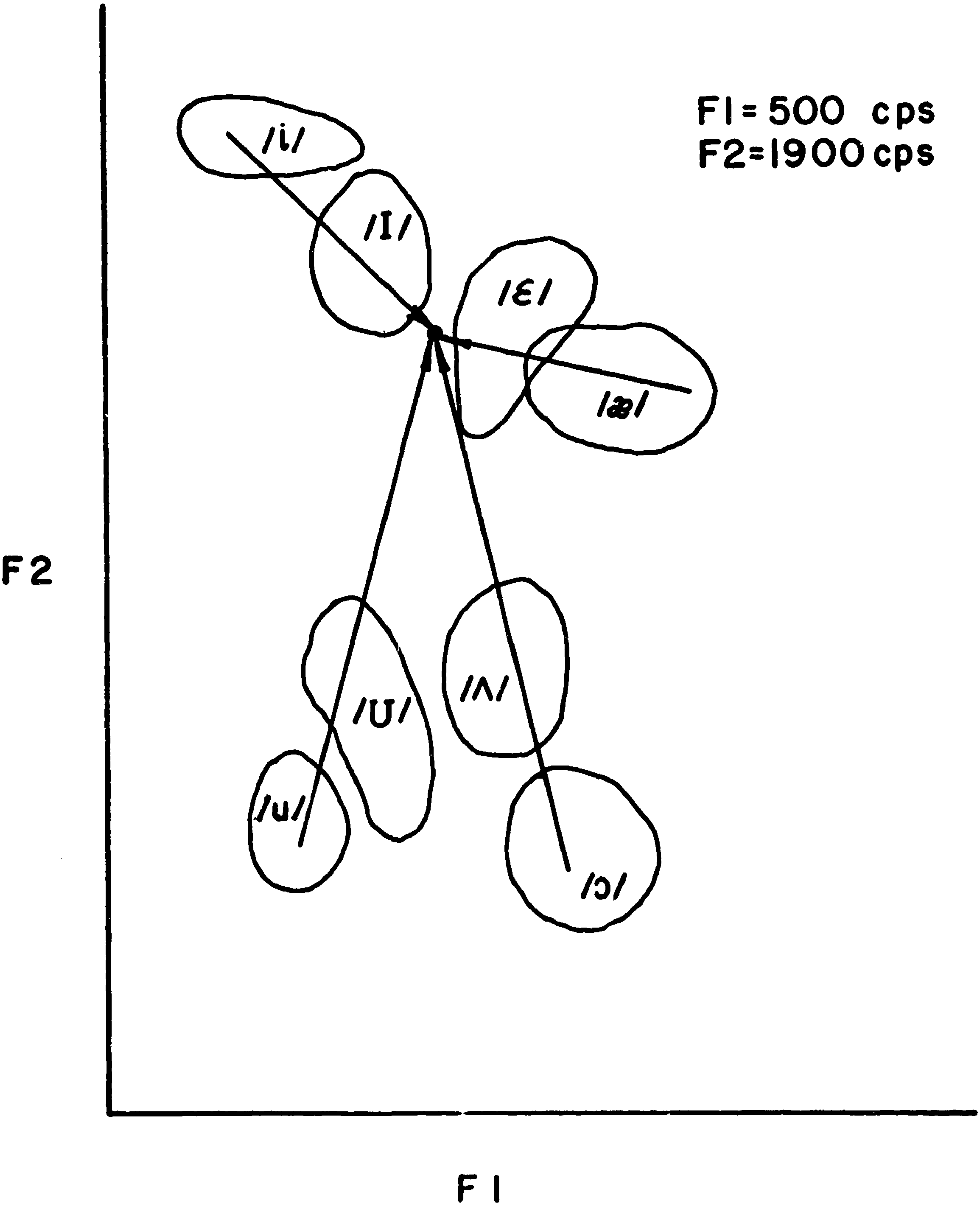

Figure III-4. Formant Regions Showing Tense-Lax Tendency (Speaker GWH).

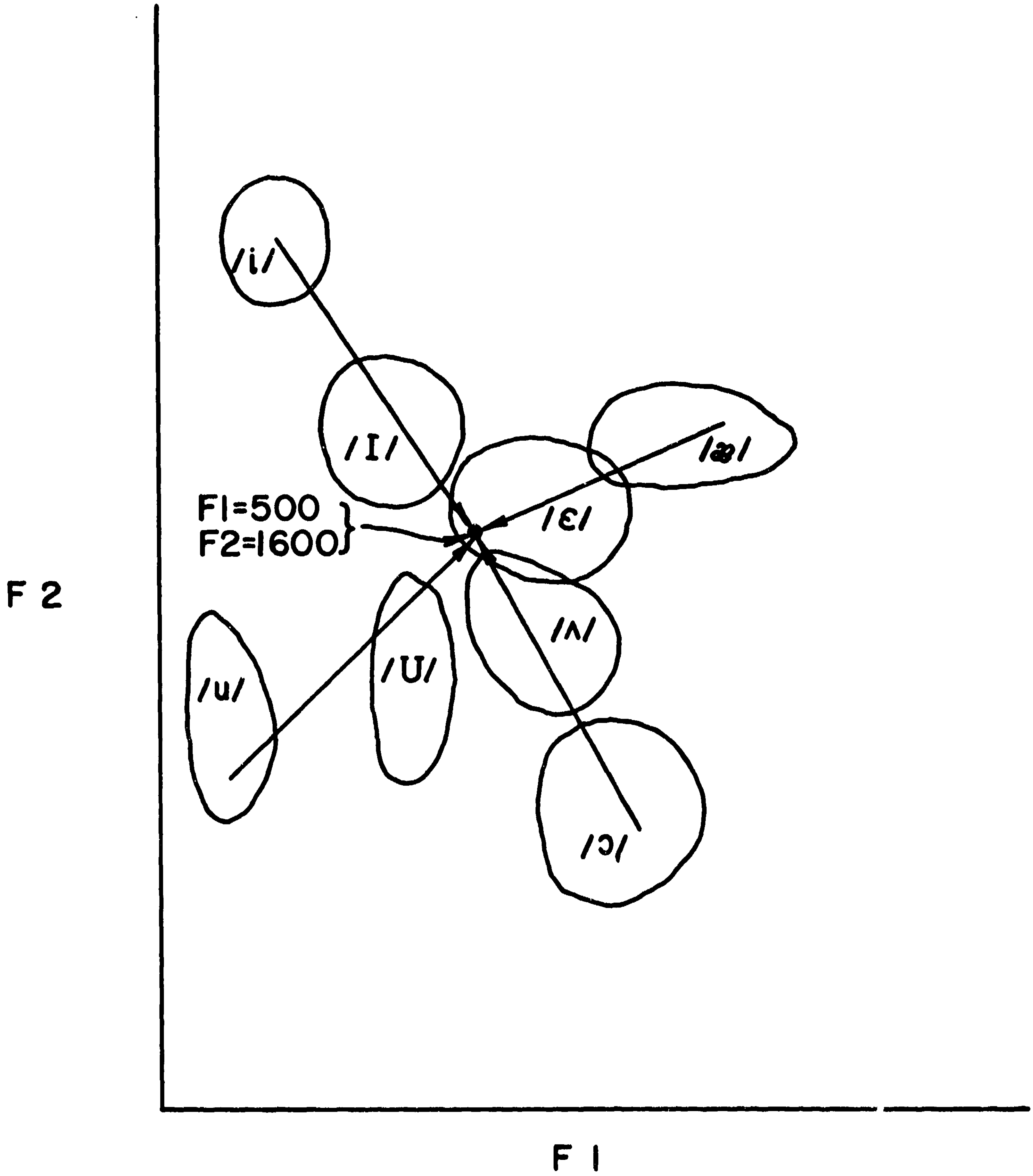

Figure 111-5. Formant Regions Showing Tense-Lax Tendency (Speaker JFH).

to be expressible in terms of more than one simultaneous physical parameter. The other features were expressed in terms of formant frequencies alone, but the tense, lax feature requires the additional and simultaneous measurement of vowel duration. This, of course, is an allowable relationship between features and physical parameter. There is no restriction implied by the distinctive feature hypothesis. This hypothesis states that the physical correlate of a distinctive feature must be expressible in terms of only one of those parameters that are commonly used for the description of signals from various sources, both physical as well as biological. For example, although cumbersome, one might define the tense/lax boundary as the surface

$$(.01T-.6f_1)\left(\frac{T}{165}\right)\left\{\frac{(f_1-5.5)^2}{9}+\frac{(f_2-17)^2}{4}\right\} - 1 \quad = \text{const.}$$

T = duration in msecs. $f_1 = \log F_1$ $f_2 = \log F_2$

This empirically determined relation is an elliptical cone in a three-dimensional F1-F2 duration parameter space. In practice, such a complicated function would not be used for machine recognition. However, as a theoretical boundary between the tense and lax vowels, it serves quite well to separate them. It also serves to point out that there is no reason to expect that the feature boundaries will be expressible by simple relation of physical parameters (frequency, time, amplitude), commonly used to describe certain complex waveforms.

The above expression for the tense/lax boundary, in spite of its awkwardness, is of theoretical value. It has been shown that by being a reasonable measure of tension for the true tense/lax pairs, this expression indicates whether the two remaining vowels /ɚ/ and /ɔ/ are tense or lax. Substituting the proper values of F1, F2, and duration into the formula indicates that both of these vowels should properly be regarded as tense, as expected.

C. Some Physical Correlates of the Consonantal Features

Because of the scope of the problem of finding the physical parameters corresponding to all the distinctive features in English, it was decided to place the emphasis of the experimental work on finding the correlates of the vowels. The vowel correlates are, however, often dependent upon the consonantal environment, and it is necessary, therefore, to provide for a partial separation of the non-sonorant consonantals. The consonantal sonorants, admittedly one of the most difficult groups, have been investigated only superficially.

1. Fricative Separation

One particularly good test of the validity of the distinctive feature description demands that the physical correlate of any given feature maintain at least the same character when applied to various phonemes, if not the same measurement values. Therefore, since a good measure of the acute (grave) features is a predominance of the higher (lower) end of the spectrum, it is to be expected that such a measure should be applicable to the fricatives, both tense and lax. This is, in fact, a relatively good measurement to perform in order to separate /s/ and /ʃ/ from /f/ and /θ/ and /ʒ/ and /z/ from /v/ and /ð/. The feature compact/diffuse for the fricatives is manifest in a greater vs. lesser spread among formants as well as a greater vs. lesser intensity. These two features were combined into one series of tests as follows:

1) A large intensity indicated either /s/ or /ʃ/ for the tense fricatives and /z/ or /ʒ/ for the lax fricatives.

2) If that peak (probably F3) lying in the range of 1290 cps to 2600 cps is greater than 2130 cps a /ʃ/ is indicated.

3) If F3 is less than 2130 cps and F4 is greater than 3500 cps it is a measure of the diffuse /s/.

4) Since /f/ and /θ/ are also diffuse, these two fricatives are not separated. They are distinguished from /s/ by the first test.

Little is known about the physical correlate of the strident/mellow feature; and since both of the phonemes /f/ and /θ/ are of very low level, we were not able to distinguish between them.

The above measures were also applied to tne lax fricatives. Therefore, the decision flow chart of Figure III-6 applies to both groups, regardless of whether the phoneme is tense or lax.

2. The Separation of Stops and Affricates

The separation of the stops and affricates poses problems which are somewhat unique when compared to the other phonemic classes. It has been recognized that although the number of features is small for this group, there is a preponuerance of physical cues and correlates representing these features. Furthermore, all of these acoustic cues do not operate simultaneously; that is, the cues may occur at different points in time, may be missing entirely, or in some cases may be representative of other than the intended phoneme and yet the phoneme is capable of recognition. Specifically, a sound may be lacking the stop burst and yet be properly recognized.

Any recognition scheme for the stops must take into account the nonspecific nature of these sounds by providing the correct recognition if the measurements produce a conclusive answer. Also it must allow for the absence of some clues or the detection of erroneous measurement values. A method which has provided good results, and which is capable of extension, is one which weighs each measurement value according to the conclusiveness of the particular physical correlate.

The physical correlates which are of value for the stops and affricates are: a) the frequency of the major energy peak of the stop burst, b) the amount of total energy in the burst c) indications of turbulence in the burst, d) the rate and frequency of the formant transitions of the vowels preceding or following

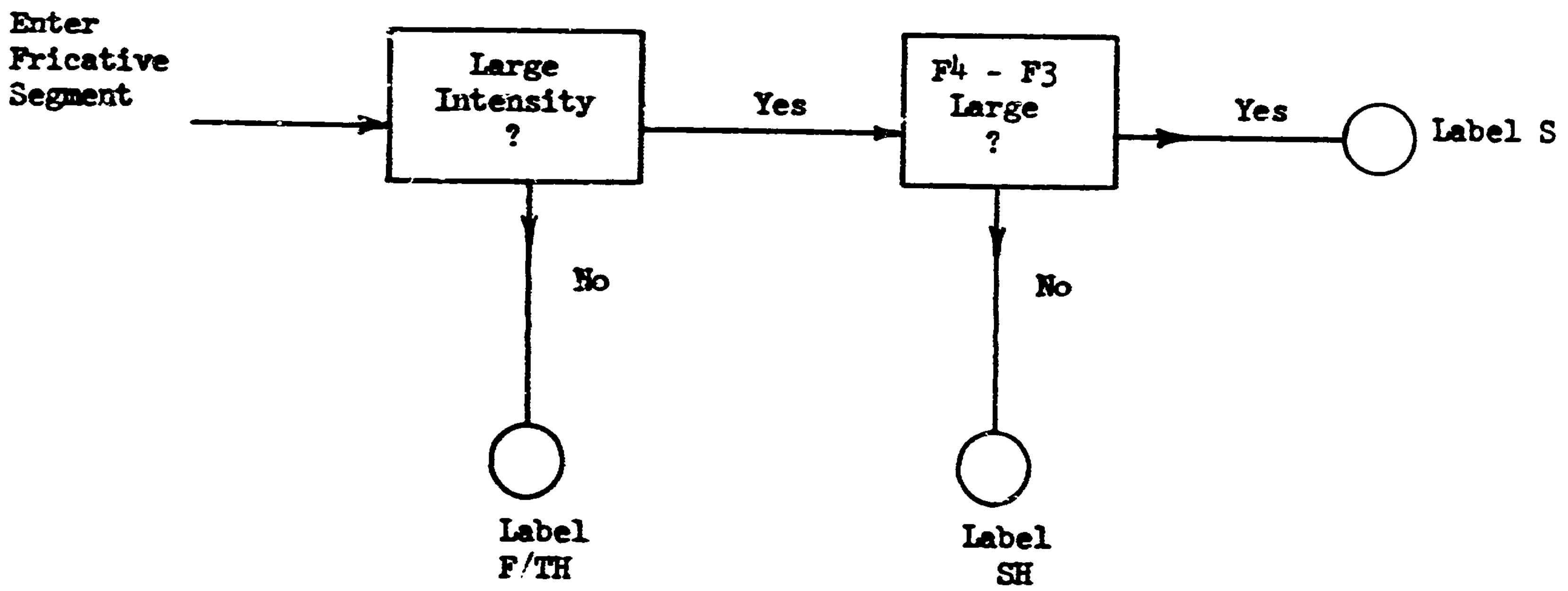

Figure III-6. Classification of Fricative Blocks.

the stops, e) the frequency and level of the stop burst poles if detectable, f) the duration of the stop burst. Only as many of these cues as are necessary are used in the recognition program.

A decision chart and the method of weighting the tense stops are shown in Figures III-7 and III-8. The various decisions have been patterned after the energy level for the compact/diffuse feature and the location of the poles or peaks for the feature acute/grave. The weighting used in each decision has been determined empirically.

D. The Addition of Simple Linguistic Constraints

A considerable increase in the accuracy of the classification is possible by the use of linguistic constraints. Results were obtained in which judgment is made on tense or lax fricatives and stops, based on the impossibility of having a tense and lax consonantal adjacent without intervening phonemes or word boundaries. For example, each fricative time segment is tentatively labeled lax or tense, according to whether voicing is detected or not. A segment is said to belong to a voiced phoneme if the energy below 350 cps exceeds a certain threshold. Although voicing and laxness are merely concomitant features in English, nevertheless, nearly all groups of time segments belonging to fricatives or stops display both voicing early in the phoneme as well as a lack of voicing toward the end, regardless of whether the speaker had intended a tense or lax sound. In many cases, if a lax fricative was intended, a clear majority of the time segments exhibit voicing. Likewise, many tense fricatives and stops have an excess of segments without voicing. However, in nearly all of the words examined which contained final fricatives, neither voicing nor lack of voicing was predominant and thus voicing could not be used as a cue for determining whether the fricative was tense or lax. It is generally known, however, that English vowels are lengthened preceding lax consonants, everything else being equal. To make use of this information in helping resolve tense/lax final consonant ambiguities, it was first necessary to determine whether the preceding vowel was tense or lax, as this vowel

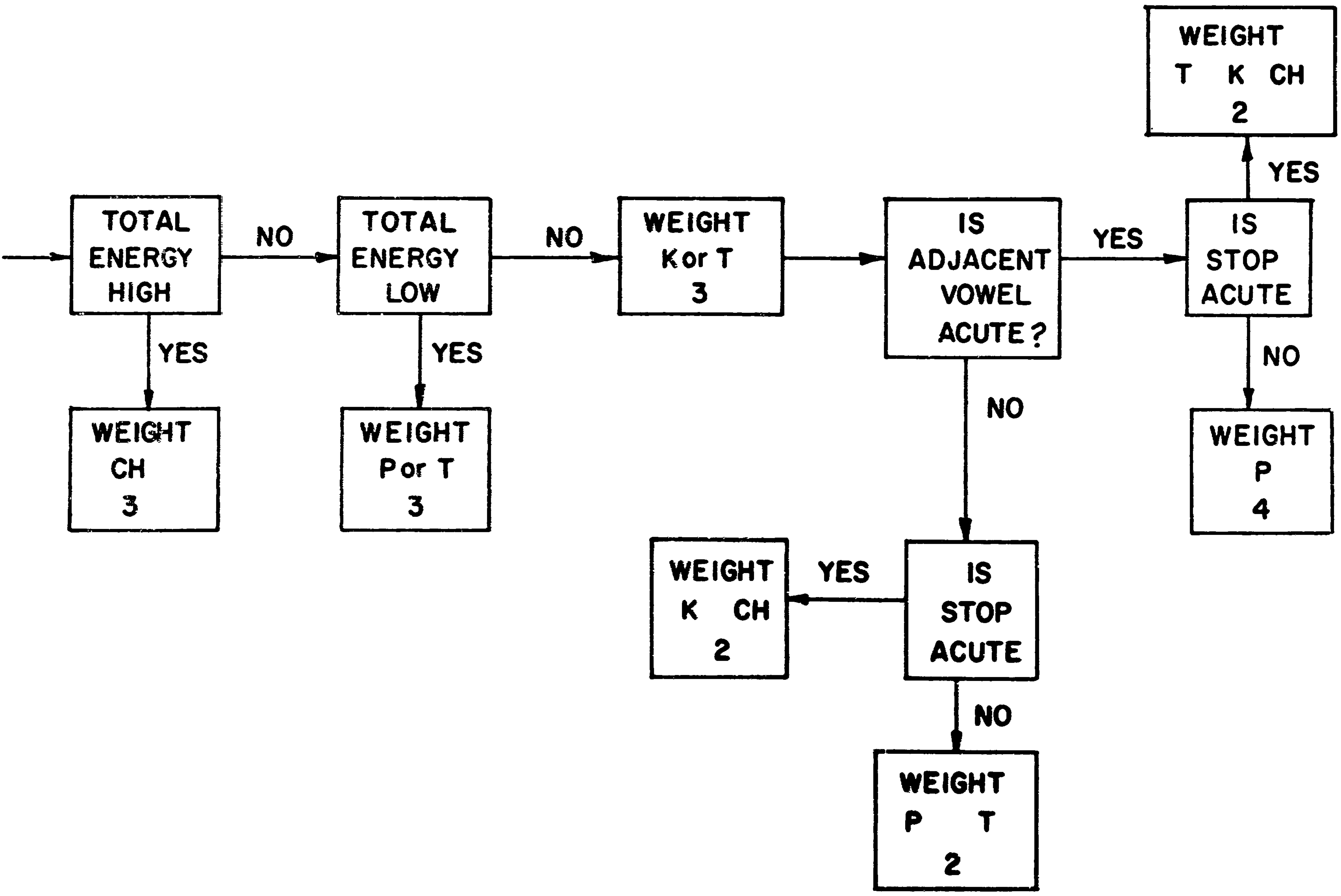

Figure III-7. Decision Chart for Stops and Affricates.

		p	t	k	tʃ
High Total Energy		0	0	3	3
Medium Total Energy		0	3	3	0
Low Total Energy		3	3	0	0
Neighboring Vowel Acute	High Frequency	0	2	2	2
	Low Frequency	4	0	0	0
Neighboring Vowel Grave	High Frequency	0	0	2	2
	Low Frequency	2	2	0	0

Figure III-8. Weighting Values for Tense Stop Decisions.

feature (as indicated earlier) is itself highly dependent upon vowel duration, everything else being equal.

The vowel is classified tense or lax, if possible, solely on the basis of the location of its first and second formants as determined by the peak picking program. Figure III-9 portrays the logical process used to decide whether the final fricative is tense or lax. If the number of unvoiced segments of a fricative block exceeds three more than the number of voiced segments, the block is labeled tense. Conversely, if the number of voiced segments exceeds three more than the number of unvoiced segments, the block is labeled lax. Otherwise the decision is made by examining the duration of the preceding vowel. If the fricative block is preceded by a vowel whose duration is greater than 250 msecs, the fricative block is labeled lax regardless of whether the vowel is tense or lax. But if the vowel is known to be lax and has a duration greater than 170 msecs, then the fricative block is also labeled lax. On the other hand, if the vowel has a duration less than 150 msecs, the final consonant is labeled tense. If the vowel is tense and has a duration less than 220 msecs, the fricative block is also labeled tense.*

Some vowels have durations which lie between the ranges just named and which do not permit consonant tense/lax resolution. The fricative blocks are then labeled lax if the number of voiced segments is equal to or greater than the number of unvoiced segments. Otherwise, the block is labeled tense.

This method of resolving tense/lax consonant ambiguities by use of preceding vowel duration was initially tested on 52 CVC words out of the master list of nonsense utterances which contained final fricatives. Half or 26 of the final

* The use of nonsense utterances allows this test to achieve high accuracy. It is apparent that normal speech will require shifting duration thresholds which depend upon the mean tempo of the speech. In addition, it will probably be necessary to recognize the presence of diphthongs and their implications for duration constraints.

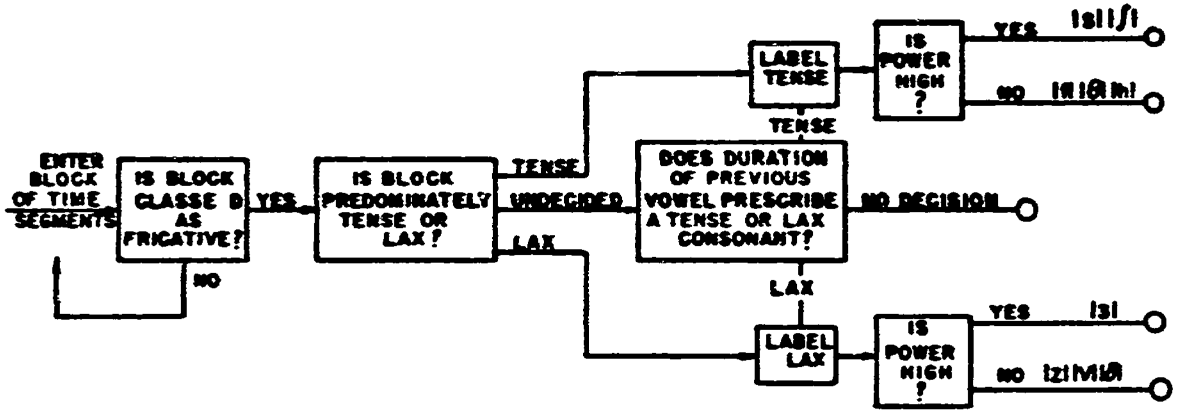
ENTER BLOCK OF TIME SEGMENTS
IS BLOCK CLASSE D AS FRICATIVE?
NO
YES
IS BLOCK PREDOMINATELY TENSE OR LAX?
TENSE
UNDECIDED
LAX
DOES DURATION OF PREVIOUS VOWEL PRESCRIBE A TENSE OR LAX CONSONANT?
TENSE
LAX
NO DECISION
LABEL TENSE
LABEL LAX
IS POWER HIGH ?
YES
/s//ʃ/
NO
/f//θ//h/
IS POWER HIGH ?
YES
/ʒ/
NO
/z//v//ð/

Figure III-9. Preliminary Fricative Classification Using Vowel Duration Infcrmation

fricatives were clearly voiced or unvoiced and were thus labeled tense or lax correctly, with the exception of one error. Of the 26 fricatives with voicing ambiguities, 18 were correctly resolved by using vowel duration information. Three were classified erroneously, and five fricatives were not decided upon on the basis of vowel duration. Three of these last five were finally classified correctly as tense or lax. Thus, 46 of the 52 or 88% of the final fricatives were labeled as tense or lax correctly. The same decision made without benefit of vowel duration resulted in an accuracy of 75%.

Duration is also known to play an important role in the vowel distinctive feature tense/lax. However, vowel duration may also be helpful in improving the detection of the vowel distinctive feature grave/acute, which primarily depends upon the location of the first formants.

Figure III-10 is a small portion of a typical F1-F2 plot, which shows the formant locations representing several examples of the four vowels /ɚ/, /æ/, /ʊ/, /ʌ/. The boundary separating the acute vowels from the grave vowels should probably lie between the dotted lines. There is considerable formant position overlap of these vowels in the neighborhood of the grave/acute boundary, even though the plot included the speech of only one individual. This, no doubt, results partially from the effect of the surrounding phonemes on the vowel formant positions. Note, however, that in this region where the most grave or acute vowels overlap, those vowels which are acute are also tense: /ɚ/ /æ/; while those vowels which are grave are also lax: /ʊ/ /ʌ/. Thus, in this region of greatest confusion, the acute vowels are of longer duration than the grave vowels - everything else being equal. Since the following consonant influences the vowel duration, this consonant must be known to be either tense or lax before an attempt is made to resolve the vowel ambiguities in this region. The computer decisions performed for vowels in this region are shown in Figure III-11. If the formant positions of the vowels lie above the upper dotted line, the vowel is labeled acute , if below the lower dotted line, the vowel is labeled grave. Vowels

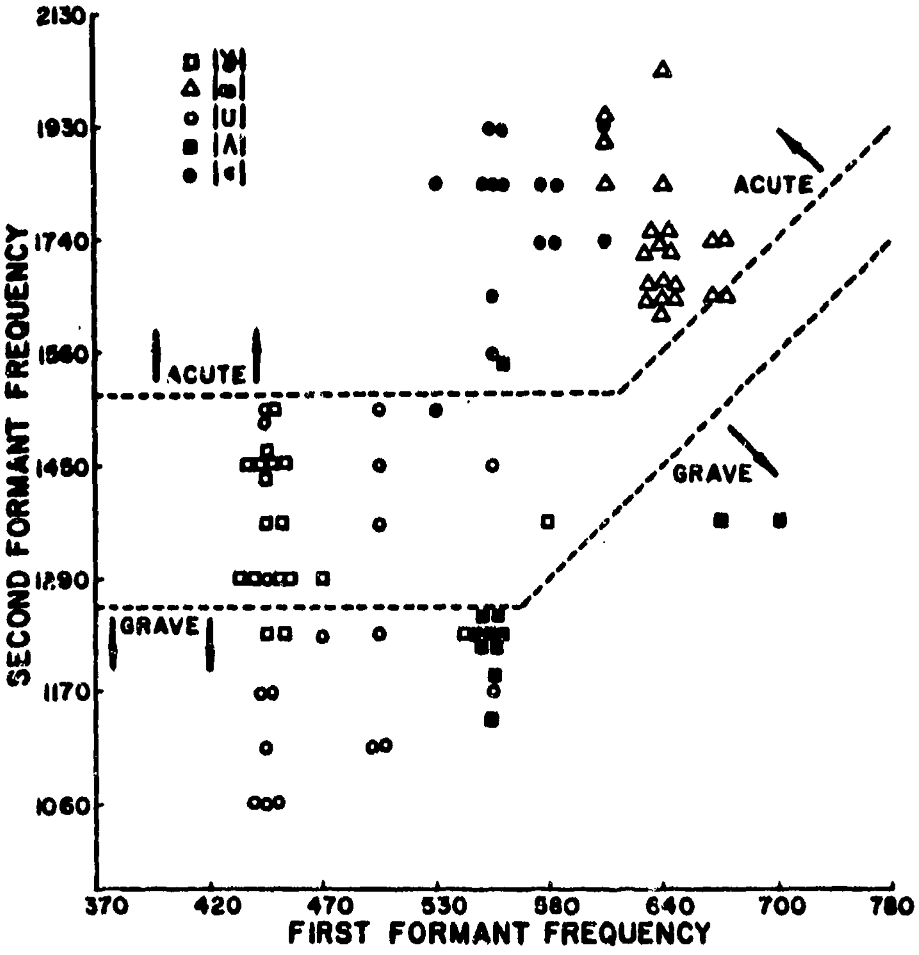

Figure III-10. Portion of F1-F2 Plane Showing Grave/Acute Boundaries

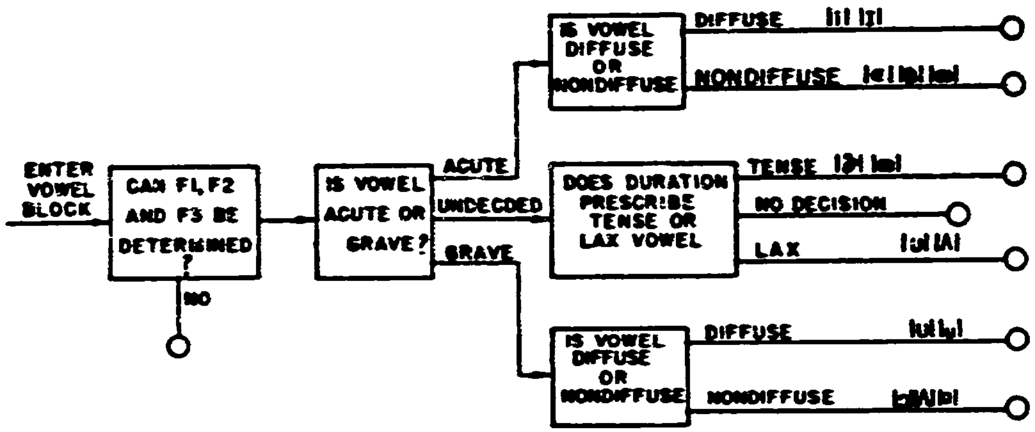

Figure III-11. Preliminary Vowel Classification Using Vowel Duration to Improve Acute/Grave Decisions.

whose formant positions lie in the region between the dotted lines, and which are followed by a tense consonant, were labeled lax, and consequently grave if the vowel duration was less than 110 msec. Conversely, they were labeled tense, and therefore acute, if the duration was greater than 170 msec. For similar vowels followed by lax consonants, the duration boundaries were chosen at 210 msec and 285 msec respectively.

E. The Complete Vowel Segment Classification

The complete decision process for each vowel time segment is portrayed in the block diagram of Figure III-12. The formant positions and vowel length determine a label for each time segment. For example, suppose a segment has spectral peaks located at 470 cps, 1430 cps and 1930 cps. Then, following the logic of the diagram, one finds that $F1<530$ cps and the segment is labeled diffuse; $F2>1170$ and the segment is labeled tense; $F2<1740$ and the segment is labeled flat; $F3<2130$ and the segment is labeled acute. The only acute, diffuse tense flat vowel is /ɚ/.

In nearly all normally pronounced vowels, the formant positions change throughout the duration of the vowel, as a result of the consonantal influence. In addition, many vowel sounds do not belong to the ten or so cardinal vowels but exhibit a diphthongal quality. Since each time segment is labeled independently from the rest, a changing series of vowel labels occurs as the output of the computer program at this stage. It is necessary at this point to decide if a cardinal vowel or a diphthong (or possibly even triphthongs) best represented the intent of the speaker. To do this, the machine must distinguish between those formant transitions which are normally a result of the surrounding consonants and those transitions which signal the presence of the more complex vowel sounds. It is necessary, therefore, to present information to the computer at the physical level which characterizes each type of formant transition. As an example of such information, it is observed that most often, diphthongs are a rapid merger of a given vowel with a more diffuse vowel following (i.e., a lowering of F1 usually occurs

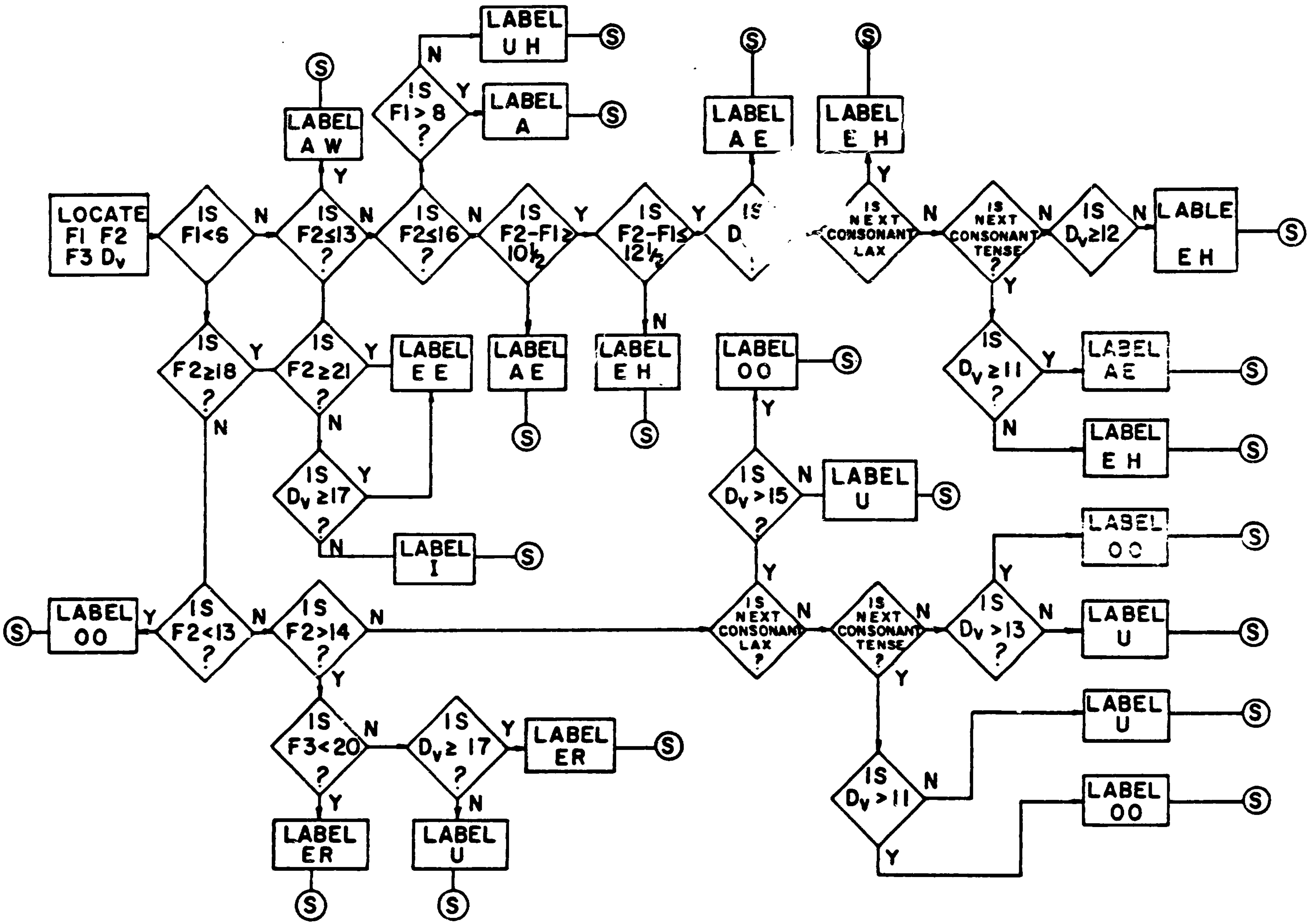

Figure III-12. Vowel Time Segment Decision Tree.
Filter Frequencies Given in Appendix III

during the progress of the diphthong), while the adjacent consontants most often affect only the second formant which may either rise or fall, depending upon whether the consonant is grave or acute. It is usual to find that the consonantal transitions are also more rapid than the diphthong (ceteris paribus). There are exceptions to these relationships, but the following logic rules provide a reasonably accurate separation of diphthongs from consonantal transitions:

1) If a given vowel classification is one of the lax vowels and lasts for at least six consecutive segments, the corresponding vowel is assumed to be present either alone or in diphthongal conjunction with another and is, therefore, an output symbol.

2) If a given vowel classification is one of the tense vowels and continues for at least eight consecutive segments, the corresponding vowel is assumed to be present and is, therefore, an output symbol.

3) If neither of these conditions has been met, then a maximum or a minimum second formant frequency is located; and if this maximum or minimum is not at the beginning or end of the vowel block, the output symbol becomes that vowel for which this extremum corresponds.

4) If there are two or more output symbols for vowels without indication of intervening consonants, then only those vowel clusters which are allowable diphthongs in English form the output symbols. Specifically, these diphthong outputs which are allowed are: /aI/, /ʌu/, /ɛI/, /aɚ/, /aU/, /ɛɚ/, /Iɚ/, /uɚ/, /ɔI/.

5) If no output has thus been achieved, the vowel block is labeled with that symbol found in the largest

number of vowel segments.

6) Certain redundant checks are then performed upon the vowel block.

As an example of the last step, suppose a certain time segment sequence of labels shows

/I/ 4 segments
/u/ 8 segments
/I/ 6 segments

Further, suppose that the speaker had intended the nonsense word of /tʃut/. The computer would ignore, because of insufficient length, the first /I/, would tentatively select /u/ and /I/ as an output sequence as a result of rules 1) and 2) above, but by rule 4) reject this as an unallowed diphthong. The program may at this point use additional information concerning the consonantal environment, or it may enter decision 3) or 5) - either one of which would produce the desired result. However, it might be of value in other circumstances to utilize additional information. In this case, the final consonant is acute, and the second formant should rise in the vowel in anticipation of the acute burst of the /t/. Therefore, the /I/ classification before the final consonant can be inferred and this classification deleted. The distinction between <u>cot</u> /kat/ and <u>kite</u> /kaIt/ comes from the greater length of the /a/ relative to /I/ in <u>cot</u> than in <u>kite</u>.

Chapter 4

EXPERIMENTAL RESULTS

A. System Input and Output

The analysis and logical procedures outlined in the preceding section are designed to produce an input/output correspondence as shown in Figure IV-1. The symbols and phonemes are not in one-to-one correspondence; for example, both /f/ and /θ/ produce a F/TH output when correctly recognized. In addition, all non-vowel sonorants which appear at the end of a word (or are immediately followed by a silence) are not individually classified but are merely labeled SON. There was no attempt to separate the nasal consonants; therefore, each of these phonemes (in medial position) is expected to receive the label M/N.

A corpus of nonsense utterences in the form /hə'C_1VC_2/ was used in this study. It was obtained by allowing V to assume the ten vowel types given in Figure IV-1 and permuting C_1 and C_2 through the rest of the phonemes listed in Figure IV-1. Talkers, whose utterances were recorded for use, spoke "standard" American (mid-western) dialect and had phonetic training. They alone determined the acoustic manifestation of the phonemic representation. Other spoken data employed to test the analysis procedures consisted of a list of 50 English monosyllabic words containing diphthongs and consonant clusters, and two English sentences.

B. Vowels

1. Confusion Matrix for the Vowels

The results of the vowel classification program can best be displayed by means of a confusion matrix, as shown in Figure IV-2. The numbers appearing along the principal diagonal are the number of correct responses. The numbers off the principal diagonal are the errors where the computer output did not agree with the intent of the speaker. A lack of correspondence with the intent of the speaker was not

IPA Symbol	IBM Output	IPA Symbol	IBM Output
i	EE	p	P
I	I	t	T
ɛ	EH	k	K
æ	AE	tʃ	CH
a	A	b	B
ʌ	UH	d	D
ɔ	AW	g	G
u	OO	dʒ	J
ʊ	U	s	S
ɚ	ER	ʃ	SH
m	M/N or SON	f	F/TH
n	M/N or SON	θ	F/TH
ŋ	SON	z	Z
l	L or SON	ʒ	ZH
w	W	v	V/TH
r	R	ð	V/TH
j	Y	h	H

Figure IV-1. Correspondence Between IPA Symbols and Computer Output Symbols. See Figure I-1 for Pronunciation Guide.

Computer Output

	i	I	ε	æ	a	ʌ	ɔ	u	ʊ	ɚ	Totals
i	23										23
I		21						2			23
ε		1	15	5		1					22
æ				23							23
a					23						23
ʌ					2	19	1				22
ɔ							23				23
u		2						20			22
ʊ		1				1	1	2	18		23
ɚ		1				1				21	23
											227

Figure IV-2. Vowel Confusion Matrix.

considered in error in those two cases for which more than half of the listeners agreed with the results of the machine. Nineteen vowel outputs are those considered in error, which results in an overall vowel accuracy of 92%. The tense vowels accounted for only four of these errors for an accuracy of 97%, and the shorter or lax vowels had fifteen errors for an accuracy of 84%.

2. Analysis of Vowel Errors by Features

The purpose of an error analysis, besides evaluating and grading the method of approach and the validity of the measurements, is to determine the areas of weakness, as made evident by a "bunching" of similar types of errors. Since the computer approach was one of tracking the distinctive features, it is instructive to portray the errors incurred in terms of the features themselves. The four features of the vowels are shown with their confusion matrices in Figure IV-3. The only important grouping of errors occurs for the tense/lax feature. However, all of the eight tense/lax errors are from vowels which were labeled lax when they should have been labeled tense, and this one-sidedness suggests that appropriate boundary changes may eliminate some of these eight errors.

3. A List of Vowel Errors and Their Probable Causes

There were 20 vowels which produced an output different from the intent of the speaker out of a total of 227 operated on by the computer program. Because of their greater length, the tense vowels had fewer errors, there being only four errors out of 137 samples in all possible consonantal environments. There were two mislabelings of the vowel /u/ which were called /I/. In both cases, the /u/ was preceded by an acute consonant and followed by a sonorant classification. This environment requires that the second formant of the vowel change rapidly downward. The vowel classification thus becomes highly sensitive to the sonorant/nonsonorant boundary location, since there is no portion of the vowel with significantly stable formant positions. In both of these errors, the sonorant boundary was signaled too far in advance of the completion of the vowel and, consequently, the formant positions had not yet arrived at the proper positions. An improvement in the

Intended Feature	Computer Feature	
	A	G
Acute	110	4
Grave	3	110

	C	D
Compact	113	1
Diffuse	5	108

	T	L
Tense	88	0
Lax	8	85

	F	P
Flat	46	0
Plain	1	45

Figure IV-3. Confusion Matrices for Vowel Features.

test for the sonorant/nonsonorant feature would remove these errors. The other two errors in the tense vowels were a wrong classification of the vowel /ɚ/. In one word the /ɚ/ was labeled a /ʌ/ because this vowel followed a nasal sound and the vowel subsequently contained a strong indication of nasalization. The peak picking program tracked this nasal peak instead of F1 and then classified the vowel as compact. The second /ɚ/ error occurred because the vowel was surrounded by two acute consonants, and although the second formant did momentarily reach the target position, it did not remain there long enough to receive the /ɚ/ label.

There were 17 errors in 90 samples of the lax vowels in all contextual environments, and the greater error rate was almost entirely due to the shortness of the vowel and the resultant failure of the second formant to reach its target value.

C. Results for the Additional Word List

In order to expand the data by which this computer program was evaluated, an additional list of 50 words was introduced. These words, spoken by the same speaker, were common English words which were expected to be troublesome in one or more ways. First of all, 20 of these 50 words contained diphthongs. All of the words contained consonant clusters either initially or at the end of the word, or both.* The departure from a given CVC form for the data words introduces errors other than a substitution of phonemes. First, there is the possibility of an output when no phoneme is present; and second, there is the possibility of no output for one or more phonemes. Although the consonant classification was poor (the correct presence of 78 out of 135 consonants was determined), there were few words for which the output did not bear a close resemblance to the intended word. For example, the word <u>beauty</u> was transcribed by the computer as DIUDI, <u>dry</u> as TAIY, clear as KEE ER,

* A complete listing of all the spoken material that was analyzed together with the computer transcription may be found in Ref. 10. Some of the words on this list were, "blind", "pure", "quick", "screen", etc.

claps as KLAES, *split* as SPWIT. In addition, many words produced a readable transcription, for example, *stunk* was transcribed as STUH $\overset{M}{N}$ K, *prove* as PROOV, *cure* as KEE OO ER, *grow* as GRAW OO, *smooth* as S $\overset{M}{N}$ OO $\overset{F}{TH}$, *flow* as $\overset{F}{TH}$ LAW OO, *sweet* as SWEET, *few* as $\overset{F}{TH}$ IUOO, *skew* as SKIUW, *gets* as GEHTS and *lacks* as LAEKS.

1. Diphthong Results

The additional word list described in the preceding section contained 20 words with diphthongal type vowels. The results obtained, although somewhat inconclusive because of the small number of samples, indicate that the vowel program, as given, will produce multiple vowel outputs which correspond closely to the intended sounds. Those diphthong outputs which contained three vowel symbols will be resolved by allowing only proper vowel pairs to remain. The outputs of 14 diphthongs out of 20 were in complete agreement with the intended words.

D. Results for Continuous Speech

The ultimate goal of all but the most trivial speech recognition devices is to transcribe the phonemes of natural and continuous speech. Although this goal is somewhat utopian in nature, it does seem appropriate to subject samples of continuous speech to the computer analysis in an attempt to exhibit those areas of the recognition problem which arise as a manifestation of the continuous nature of natural speech. Two short meaningless sentences were processed by the computer program: "Yowie, ma knew me while I rang" and "A lamb wants your material." These sentences (a), along with a phonemic transcription (b) and the results of the vowel classification program (c), are given below:

(a) "Yowie, ma knew me while I rang"
(b) /jowi ma nu mi waIl aI ræŋ/
(c) /Iɔi/ /ʌ/ /ʊ/ /I/ /a/ /a/ /ɛ/

(a) "A lamb wants your material"
(b) /ɛI læm wants jɔɚ mʌtIɚiəl/
(c) /i/ /æ/ /ʌ/ /Iu/ /ʌ/ /ɚiɔ/

E. Comparison of the Results with Listening Tests

Natural speech is a means of communication between two or more individuals and the effectiveness of this communication can be judged only in terms of the response and behavior of the listeners. This study is concerned with the recognition of this natural speech and, therefore, the judgment of the "correctness" of the machine transcription must be based on the agreement of a panel of listeners of the same language group (if agreement is possible). With this necessity in mind, a series of listening tests was given in which a number of first year phonetics students were presented with 100 randomly chosen nonsense words from the original hə'CVC word list. The conditions of the test were designed so that the group of listeners was presented the same stimulus as the computer under nearly the same conditions. All 25 listeners were asked to listen to each nonsense word twice and to transcribe what they heard as accurately as possible. The nonsense words used for these tests were dubbed directly from the tapes used to provide the input for the computer, and therefore exactly the same waveforms were presented to both computer and panel.

There was unanimous agreement as to the intent of the speaker for 30% of the nonsense words. There was complete agreement for 60% of the words by 80% of the listeners (20 out of 25). This last figure includes listener agreements which did not correspond to the intent of the speaker. There were three cases in which at least 80% of the listeners produced an agreement which differed from the intent of the speaker. Sixteen words produced almost complete confusion, as only 60% or less of the listeners gave the same response.

The results differ somewhat, if only the vowel responses are considered. For vowel recognition only, 80% or more of the listeners agreed on 83% of the vowels. Eight vowels produced confusion, with not more than 60% of the listeners in agreement. The most common of these eight confusions was /ʌ/-/ʊ/, which accounted for six. (Complete listing of listener judgments is given in Ref. 10.)

It could be argued that under the conditions of the experiment (nonsense words spoken twice to a group of 25 first year phonetic students), primary recognition by a panel of listeners was comparable to the accuracy obtained with the computer program (around 91%). However, overall accuracy alone is not a sufficient measure of the success of the device. The boundaries and thresholds of the decisions for the computer should produce results, *including the errors*, which were obtained by a majority of the listening panel. Sounds which result in confusion in the panel of listeners should produce similar mistakes or confusions in the machine.

Of the eight vowel confusions resulting from the listening tests, four of the vowels were also classified incorrectly by the computer. Of these four, two resulted in a response by the computer which was the same response of the majority of the listeners and yet differed from the intent of the speaker. For example, the intended word /jɪd/ resulted in a computer classification of /jʊd/, and nearly half of the panel of listeners preferred the /jʊd/ transcription over the /jɪd/. Similar results were obtained for the pair /sʊm/-/sʌm/.

Thirteen of the 100 words presented to the listeners produced a computer output which differed from the intent of the speaker. However, in only four of these 13 words did 80% or more of the listeners agree. On the other hand, 17 of the 100 words resulted in less than 80% agreement, and eight words resulted in less than half the listeners in agreement as to the vowel present. Thus, depending upon the confidence level chosen, the machine did slightly better or slightly worse than a panel of listeners.

The word list of hə'CVC words used in this study was contrived in an artificial manner in order to insure that all CV and VC combinations would occur. It was expected, then, that some of the words on the list would be identical to meaningful English words. Of the list of 100 of these words presented to the panel of listeners, 22 were meaningful - for example, *ping*, *check*, *Bert*, *but*, *bat*, *sub*, *bought*, *bit*, *bin*, *pit*, *jet*, *sawed*, *heed*, *full*, *cad*, *jot*, *curb*, *doom*, *fill*, *reet*, *yes* and *thud*. Of these 22 words, *only one* was missed by more than 20 of the listeners. Furthermore, the most common error in five other words was the substitution of the correct words by a wrong but meaningful English word - for example, /fut/-/fʊt/, /bʊŋ/-/bʌŋ/, /dʊn/-/dʌn/, /ʒus/-/dʒus/ and /sʊm/-/sʌm/. The results suggest that listeners make use of knowledge of the phonemic structure of possible English words.even though the words are spoken in isolation, and the listeners expected nonsense words.

F. Vowel Formant Data of an Additional Speaker

The same list of nonsense monosyllables was spoken by another person, recorded and processed in the same manner as the words of speaker one. The formant location of this speaker are displayed on the F1-F2 plot of Figure IV-4. Only the "true" tense/lax pairs have been included. Several interesting differences can be seen. As was expected, the vowel regions have shifted somewhat relative to the regions of speaker one. It is apparent, by comparing with Figure III-3, that there can be no single boundary which will separate these vowels equally well for each speaker. It seems unlikely that such boundaries exist. However, adaptation of the recogniti ı device, which takes on the form of shifting boundaries in this case, will allow recognition accuracies to be nearly equal and high. Some of these boundary shifts can be determined by an examination of the formant positions of the additional speaker through the techniques used in determining the boundaries for speaker one. For example, it appears that a simple and accurate grave/acute boundary may be a line of constant F2 at about 1600 cps. It will also be noted that there is very little overlap among the tense/lax pairs relative to the amount recorded in the formant plots of speaker one. An oval, very similar to the one described in chapter two, would serve very well to separate the tense/lax pairs. The flat/plain feature would again be simple and of the same nature as for speaker one. However, the boundary for the compact/diffuse feature no longer appears to be the simple "F1 equals a constant" relationship, although an accurate boundary would be very close to a vertical straight line. Even though the compact/diffuse boundar

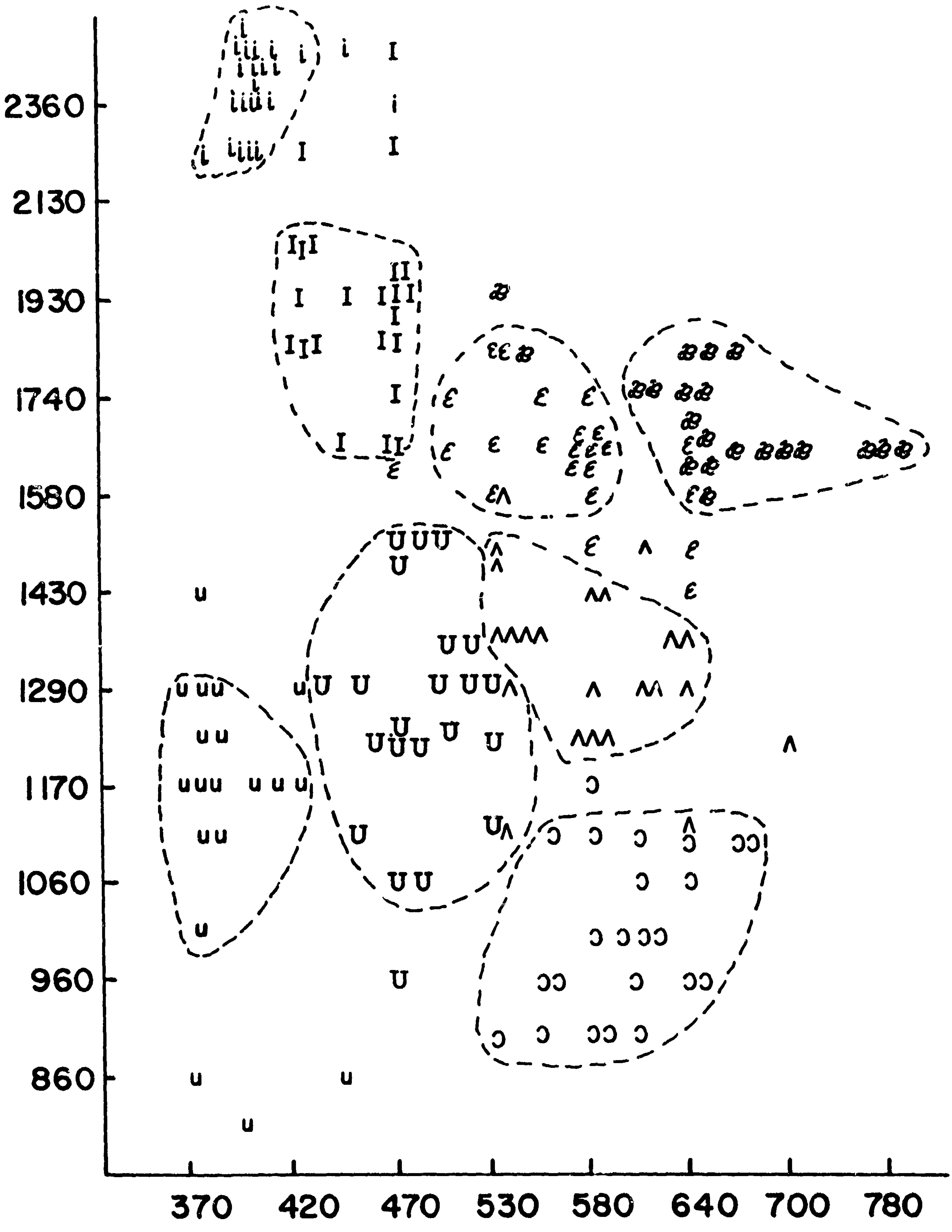

Figure IV-4. Formant Locations For Speaker JFH.

no longer takes on such a simple form, nevertheless, good separation is possible.

Very simple shifts of the boundaries of speaker one result in good, although perhaps not the best, boundaries of separation for speaker two. Thus, by increasing every point of the grave/acute boundary of speaker one by approximately 180 cps, the new locus serves as a reasonable boundary for speaker two. Likewise, the new compact/diffuse boundary can be obtained by shifting the old boundary to the right on speaker one. The approximate location of the intersection of the three boundaries - acute/grave, flat/plain, and compact/diffuse - may be given as F1 = 500 cps and F2 = 1300 cps, while the focus of the tense/lax tendency appears to merge at F1 = 500 cps and F2 = 1900 cps. Thus, for speaker one these two hypothetical points were widely separated, while for speaker two the points were almost identical.

Summarizing the results of speaker two and comparing them with the formant positions of speaker one, it can be said that the vowel region shifts rule out the possibility of retaining the same vowel decision boundaries for all speakers and yet provide good separation. On the other hand, simple shifts or translations of the boundaries of speaker one result in usable boundaries for speaker two.

Chapter 5

SUMMARY AND CONCLUSIONS

This report has dealt with the measurement or analysis of the speech waveform at a very low level of complexity. Termed here "primary recognition", we have taken this to mean essentially a time-segment by time-segment classification of the data into phonemic, or at least phonetic, categories. We have been concerned with measurement more than theory. However, the underlying link to linguistic significance is the framework of distinctive features; and it is at least a part of the physical manifestation of this theory that we seek.

The framework of distinctive features is much more fruitful in producing rules and structure for speech production than for acoustic analysis. It is clearly at levels of complexity just above the acoustic analysis of waveform that the theory of distinctive features is dominant. Yet, primary recognition, or acoustic feature tracking, remains an indispensible part of the talker-listener sequence no matter what further processing is postulated. Simple schemes, such as the one described here, must be the starting point for many more complicated systems.

The central questions of mechanical processing studies at the acoustic level are how much information is in the speech waveform and how best to remove it. There is no longer any doubt that linear-programmed manipulation of the waveform or its spectrum will not produce any recognition scheme capable of approaching listener capabilities. There are many reasons for this, the limitations imposed by segmentation being the most important. Much information is lost by assuming that independent segments are bounded by points in the waveform at which one or more acoustic features change. Of course, mutual dependencies among segments extend over various acoustic and linguistic boundaries. As a first step to improving machine recognition, a stress analysis must be included which extends over complete sentences and involves at least the parameters of intensity and fundamental frequency.

Results of this study, as far as "percent correct" go, show about the same achievement as many recognition procedures. This includes listener performance, given the identical stimulus as was the computer. In the experiments reported here, both nonsense words and gated vowel segments were presented to both man and machine. The performances were quite comparable. It appears reasonable to conclude that the recognition figures achieved (80-90%) approaches the upper limit of direct segmental processing. It is our opinion that further refinements, such as analysis-by-synthesis, will not materially improve the situation - at least not until the full theory of distinctive feature phonemics can be introduced into speech analysis. This presupposes a programmable grammar.

Perhaps the singular feature of this report is the insistence that the single-speaker approach - that is, the study of various speakers "in series" rather than "in parallel" - is not a retreat from generality, but only a rearrangement of methodology. In fact, our data has pointed up some general relationships or invariances, particularly in regards to F1-F2-plane boundaries.

At any stage in this study of the recognition problem, what remains unsolved may be simply identified from that body of a priori information about each talker necessary as an input to the processing logic. As with any scheme based on distinctive feature analysis, a partial solution may also have practical use in voice-operated devices. Compared to schemes developed from studies of many talkers, those requiring some specification of talker characteristics separate more phonemes, but from a smaller variety of voice types.

Finally, the exact specifications of the input or pre-processing equipment appeared not to be particularly important. An exception to this observation lies in the need this study made evident to us for more careful attention to the low-frequency speech spectrum. Accurate first-formant tracking, low-frequency stress cues, and the low-frequency cues associated with non-vowel sonorants argue for frequency analysis (a set of filters) of good precision below 500 cps.

Appendix I

A DISTINCTIVE FEATURE LOGIC FOR THE RECOGNITION OF THE SPEECH OF A SINGLE SPEAKER*

Considerable variations in the formant locations of sustained English vowels have been observed which arise as a result of individual speaker differences and the consonantal context in which the vowel is embedded. Quite frequently utterances of differently classified vowels exhibit formant position overlap when the measured frequencies of the first formant are plotted against the measured second formant frequencies. Barney and Peterson reported results of listening tests for different vowels in the same context which show some formant position overlap even though the vowels were unanimously classified by the listeners. However, plots of first formant versus second formant of vowels spoken by a single speaker in the same context show no overlap in formant position. Barney and Peterson conclude that not only are formant variations smaller for one speaker than for several, but also that there are statistically significant differences between speakers. This is also suggested by Fant who pointed out that female speakers average about 15% greater difference between the first two formants as a result of a shorter vocal tract. The smaller formant variations for a single speaker are most likely due to certain unalterable parameters of the vocal tract as well as to speaking habits arising from the talker's background and dialect. Kersta's findings reported recently before the Acoustical Society of America strongly support this inter-speaker variability and intra-speaker stability.

Allowing the possibility of statistically significant variations in speech waveform parameters as a result of speaker differences, then the incorporation of prior information concerning a given speaker into the recognition scheme could

* Paper delivered before the 65th meeting of the Acoustical Society of America in New York City, June 1963.

reduce the necessary measurements and logic while maintaining high recognition accuracy. The problem remains to determine the minimum amount of prior information necessary to optimally adjust the recognition logic to a new speaker. An approach adopted for testing the desirability of using prior speaker information has been undertaken by designing a recognition scheme initially around a single speaker. If, then, the speech data from another speaker is introduced into the same recognition program, the required changes in thresholds and logic necessary to adapt the program to the new speaker are directly related to the amount of prior information that is needed by the particular recognition scheme. In effect, then, a single-speaker-adjusted recognition program can be the means by which only essential multi-speaker parameter statistics are gathered. A natural result of this approach is the exclusion of all speaker differences which are not meaningful from the recognition point of view. Hopefully, examination of the set of prior information required for any or all speakers may be amenable to generalization which would reduce the needed amount of prior information.

In the present study, the measurements and classifications were carried out by programming an IBM 7090 computer to operate on speech data obtained by sampling 60 times a second the outputs of 35 bandpass filters. A list of 250 nonsense consonant-vowel-consonant words produced by one speaker were first examined and later augmented by several bisyllabic words and short sentences.

The initial classification of the speech data was performed by categorizing each 16 2/3 msec time sample into one of eight classes on the basis of its own spectral characteristics, without the use of any information from surrounding time intervals. An accuracy of better than 90% was obtained at this level without any contextual constraints being applied since only those distinctive features whose acoustic properties reside in the phoneme itself were being detected.

Slide one (see Figure III-1, page 27) is a simplified description of the measurements and decisions performed on each 16 2/3 msec time interval. The block in the upper left indicates a test determining whether the given time segment being examined is representative of speech sound or a silence. The sound segments are then classified as either sonorant or nonsonorant. This is done, as in the next three blocks along the left side, by first eliminating segments belonging to stop and fricative classes. A segment is classed as nonsonorant if it lacks voicing. These unvoiced segments are further classified as continuant or interrupted depending on whether or not the segment is preceded by a silence. Further segments, representing lax fricatives, are classified as nonsonorant by detecting the presence of relatively large amounts of energy at the higher frequencies, that is, by detecting the presence of turbulence. Separating nonsonorant time segments in this manner results in categorizing the stops and fricatives into four separate groups.

The time segments labeled sonorant are further classified as belonging to either consonantal or nonconsonantal phonemes. This classification, dividing the vowel segments from the non-vowel sonorants, is based on the property that the non-vowel sonorants are weaker sounds than the neighboring vowels. A downward change in spectral energy with time is characteristic of a vowel-non-vowel boundary, while an increase in energy signals a non-vowel/vowel boundary.

Following this initial segmentation of the speech data, an approach to formant tracking in the sonorant phonemes was undertaken by locating the first three spectral peaks in the appropriate formant regions.

Slide two (see Figure III-2, page 28) shows the results of the peak picking as white dots superimposed on a standard spectrogram. The sentence used in this example was "Yowie, ma knew me while I rang", which was chosen to contain only sonorant phonemes. The peak picking approach that was used successfully tracked the first three formants in most of the vowels and many of the non-vowel sonorants with the notable exception of failing to track the zero-influenced second formants of the nasal consonants. It is apparent that more restrictive peak constraints need to be applied to the non-vowel sonorants.

The initial time segment-by-time segment classification, without benefit of class smoothing or linguistic constraints, achieves an accuracy of about 90%. That is, 90% of all time segments are correctly labeled as members of the phonemes intended by the speaker. Phoneme classification accuracy is somewhat less since measurement errors tend to be concentrated in the weak and short phonemes such as /h/, /f/, /g/, /θ/, and /v/.

A considerable increase in the accuracy of the classification at this level is possible by the use of linguistic constraints. Results have been obtained in which judgment is made on tense or lax fricatives based on the impossibility of having a tense and lax fricative adjacent without intervening phonemes or word boundaries. The fricative time segments have been separated tentatively by labeling lax or tense those segments which have or do not have voicing present respectively. Although voicing and laxness are concomitant features in English, nevertheless nearly all groups of time segments belonging to fricatives display both voicing early in the sound as well as a lack of voicing toward the end, regardless of whether the speaker had intended a tense or lax fricative. In many cases, if a lax fricative was intended a clear majority of the time segments exhibit voicing. Likewise, many tense fricatives have an excess of segments without voicing. However,in nearly half of the words examined which contained final fricatives, neither voicing nor lack of voicing was predominant and thus voicing could not be used as a cue for determining whether the fricative was tense or lax. It is well known, however, that vowels are generally lengthened preceding lax consonants, everything else being equal. To make use of this information in helping resolve tense/lax final consonant ambiguities, it was first necessary to determine whether the preceding vowel was tense or lax as this vowel feature is itself highly dependent upon vowel duration, everything else being equal.

The vowel is classified tense or lax, if possible, solely on the basis of the location of its first and second formants as determined by the peak picking program. Slide three (Figure III-9, page 42) portrays the logical process used to decide whether the final fricative is tense or lax. If the unvoiced segments of a fricative block exceeds three more than the number of voiced segments , the block is labeled tense. Conversely, if the number of voiced segments exceeds three more than the number of unvoiced segments, the block is labeled lax. Otherwise, the decision is made by examining the duration of the preceding vowel. If the fricative block is preceded by a vowel whose duration is greater than 250 msecs, the fricative block is labeled lax regardless of whether the vowel is tense or lax. But if the vowel is known to be lax and has a duration greater than 170 msecs, then the fricative block is also labeled lax. On the other hand, if the vowel has a duration less than 150 msecs, the final consonant is labeled tense. If the vowel is tense and has a duration less than 220 msecs, the fricative block is also labeled tense.

Some vowels have durations which lie between the ranges just named and which do not permit consonant tense/lax resolution. The fricative blocks are then labeled lax if the number of voiced segments is equal to or greater than the number of unvoiced segments. Otherwise the block is labeled tense.

This method of resolving tense/lax consonant ambiguities by use of preceding vowel duration was tested on 52 CVC words with final fricatives. Half, or 26 of the final fricatives, were clearly voiced or unvoiced and were thus labeled tense or lax correctly with the exception of one error. Of the 26 fricatives with voicing ambiguities, 18 were correctly resolved by using vowel duration information. Three were classed erroneously and five fricatives were not decided upon on the basis of vowel duration. Three of these last five were finally classified correctly as tense or lax. Thus 46 of the 52, or 88% of the final fricatives, were labeled as tense or lax correctly. The same decision made without benefit of vowel duration resulted in an accuracy of 75%.

Duration is also known to play an important role in the vowel distinctive feature tense/lax. However, vowel duration may also be helpful in improving the detection of the vowel distinctive feature grave versus acute, which primarily depends upon the location of the first two formants.

Slide four (see Figure III-10, page 44) is a small portion of a typical F1-F2 plot showing the five vowels /ɛ/ /æ/ /ɚ/ /ʌ/ /ʊ/. The boundary separating the acute vowels from the grave vowels should lie probably between the dotted lines. There is considerable form nt position overlap of these vowels in the neighborhood of the grave/acute boundary even though the plot includes the speech of only one individual. This, no doubt, results from the effect of the surrounding phonemes on the vowel formant positions. Note however, that in this region where the most grave and acute vowels overlap, those vowels which are acute are also tense, namely, /ɚ/ /æ/, while those vowels which are grave are also lax, /ʌ/ /ʊ/. Thus, in this region of greatest confusion, the acute vowels are of longer duration than the grave vowels, everything else being equal. Since the following consonant influences the vowel duration, this consonant must be known to be either tense or lax before attempting to resolve the vowel ambiguities in this region. The computer decisions performed for vowels in this region are shown in slide five, (see Figure III-11, page 45). If the formant positions of the vowels lie above the upper dotted line, the vowel is labeled acute. If below the lower dotted line, the vowel is labeled grave. Vowels whose formant positions lie in the region between the dotted lines, and which are followed by a tense consonant, were labeled lax and consequently grave if the vowel duration was less than 110 msecs, or tense and therefore acute if the duration was greater than 170 msecs. For similar vowels followed by lax consonants, the duration boundaries were chosen at 210 msecs and 285 respectively.

Of the 23 words whose vowel formant positions fell in this region, six were followed by sonorant phonemes and were not considered further. Of the 17 remaining vowels, 15 were correctly labeled grave or acute by this method.

Indications are that vowel duration information can be used successfully for improving the decision accuracy for several distinctive features. However, many problems remain to be further investigated. The changes in the duration thresholds necessary to take into account multispeaker differences or variations in speaker tempo have not been examined. Also, if the class of vowel plus duration helps determine the class of consonant, and class of consonant plus vowel duration helps determine class of vowel , occasions will arise where neither can be completely determined. In some cases the vowel is sufficiently long to prescribe both a tense vowel followed by a lax fricative, or sufficiently short to prescribe a lax vowel followed by a tense fricative even though formant positions or measures of voicing are not sufficient alone to allow an accurate decision to be made. A complete lack of measurable information about tense/lax pairs occurs surprisingly rarely.

Appendix II

ON THE INTRODUCTION OF PRIOR INFORMATION INTO A COMPUTER-PROGRAMMED SPEECH RECOGNITION SCHEME[*]

All practical speech recognition schemes introduce a certain amount of a priori information concerning the input set of acoustic patterns. Furthermore, it is apparent that all workers in this field, other than those claiming to study machine learning or artificial intelligence, strive to incorporate as much of this information into their recognition schemes or logic as they possibly can. One interpretation of this is simply that many research workers in speech recognition study thoroughly the fundamentals of acoustics, phonetics, motor articulation, linguistics, etc., <u>before</u> programming recognition logic and processing large amounts of data. Another is that some also back off a bit from the general problem of all phonemes and/or all speakers, and thus limit their input set in an artificial, but hopefully not random nor linguistically irrelevant, way.

There now seems to be general agreement that the phoneme and distinctive feature framework offers the most general and complete analysis of linguistic behavior. However, it also offers a minimum of suggestions on how to go about instrumenting practical measurement procedures. Consequently, we have seen the re-working of language structure to better fit present-day knowledge and techniques. More useful basic units are proposed from time to time such as the phonoid, syllable or even word. The existence of physically difficult distinctive features, such as tense-lax, etc., is denied. Although information about specific utterances given by the general linguistic framework adopted is small, the consequences upon the linguistic relevance of the derived recognition scheme are tremendous.

[*] Paper delivered before the 65th meeting of the Acoustical Society of America, New York City, June 1963.

Much more useful results have been made in recent years by improving our understanding of articulatory constraints. The work of Fant, Stevens, House, Heinz and many others enables us to incorporate into recognition schemes vital information about the signal source, at little cost in generality.

A great deal more information about the signal source, or about a special conception thereof, may be incorporated but always with loss in generality of results. Finite sets of stored patterns of easily measured parameters or severely limited vocabularies of complex units, such as words, may be the starting point for a recognition scheme. Thus, the designer and the scheme assume much and are asked to do little.

Somewhere, in between the extremes, perhaps, is the introduction of some information concerning the <u>gross</u> characteristics of the source which is the talker, himself. We are running a series of experiments and programmed recognition schemes to attempt an evaluation of the role of speaker differences in obscuring the acoustic correlates of the distinctive features. I will have time here to present only some results of a more detailed preliminary investigation into one facet of this problem.

I have chosen to show some of the work connected with a particularly difficult vowel distinction - that between /ɛ/ and /æ/. A distinctive feature analysis shows this difference to reside either in tense versus lax or in compact versus non-compact.

Three measurements were made:

1) The frequency positions of the first and second formants, determined by a computer program operating on digitalized spectral data.
2) The vowel duration, determined by observation of an oscilloscope display of the speech waveform.
3) Forced choice judgments of a panel of listeners, presented with a gated central portion of the vowel.

A plot of the F1, F2 computer-determined locations for 69 utterances of each of these vowels by three speakers in /hə'CVC/ context is shown in Figure II-a In this list the vowel was preceded and followed by all possible English consonants. Note not only the general trend of separation but also the considerable overlap. It is difficult to envision a criterion based on a function of F1 and F2 which would separate these two classes. Furthermore, the addition of more speakers would proably increase the overlap.

A tabulation of the vowel durations is given in Figure II-b. The subscript "v" designates those vowel utterances followed by a voiced or lax consonant. The traditional short-long relationship between lax versus tense vowels and lengthening of the preceding vowel by a lax or voiced consonant is apparent. The middle two columns illustrate a cancelling of these effects.

Twenty-four subjects for the listening test were presented with 100 msec portions of the vowels, whose formants and durations were shown on these figures, to judge as /æ/ or /ɛ/. Figure II-b indicates that length may be a cue for this distinction; this test sought to assess the role of formant positions. Although the subjects were naive and the sounds unnatural to them, the judgments of 20 or more of them were in concurrence for 42% of the items on the test, and 16 or more of them concurred in their judgments of 68% of the items. Their agreement, however, with the intent of the speaker will be shown on the next few figures. In general, they were biased in judgment towards /ɛ/, probably due to the consistent short length of the stimuli.

As an interesting sidelight, a second test was given them immediately following the first. Fifty utterances of just one speaker (myself), which had appeared on the first test, were re-randomized and presented to the subjects who were informed as to who was speaking the test items. I pronounced each vowel separately and in context for them. The uniformity of their judgments went up from, for example, the previous 42% (20 or more subjects) to 64% of the items. However, on the first test they differed in their judgments from the intent of the speaker on three of my speech items, whereas on the second test they differed in seven.

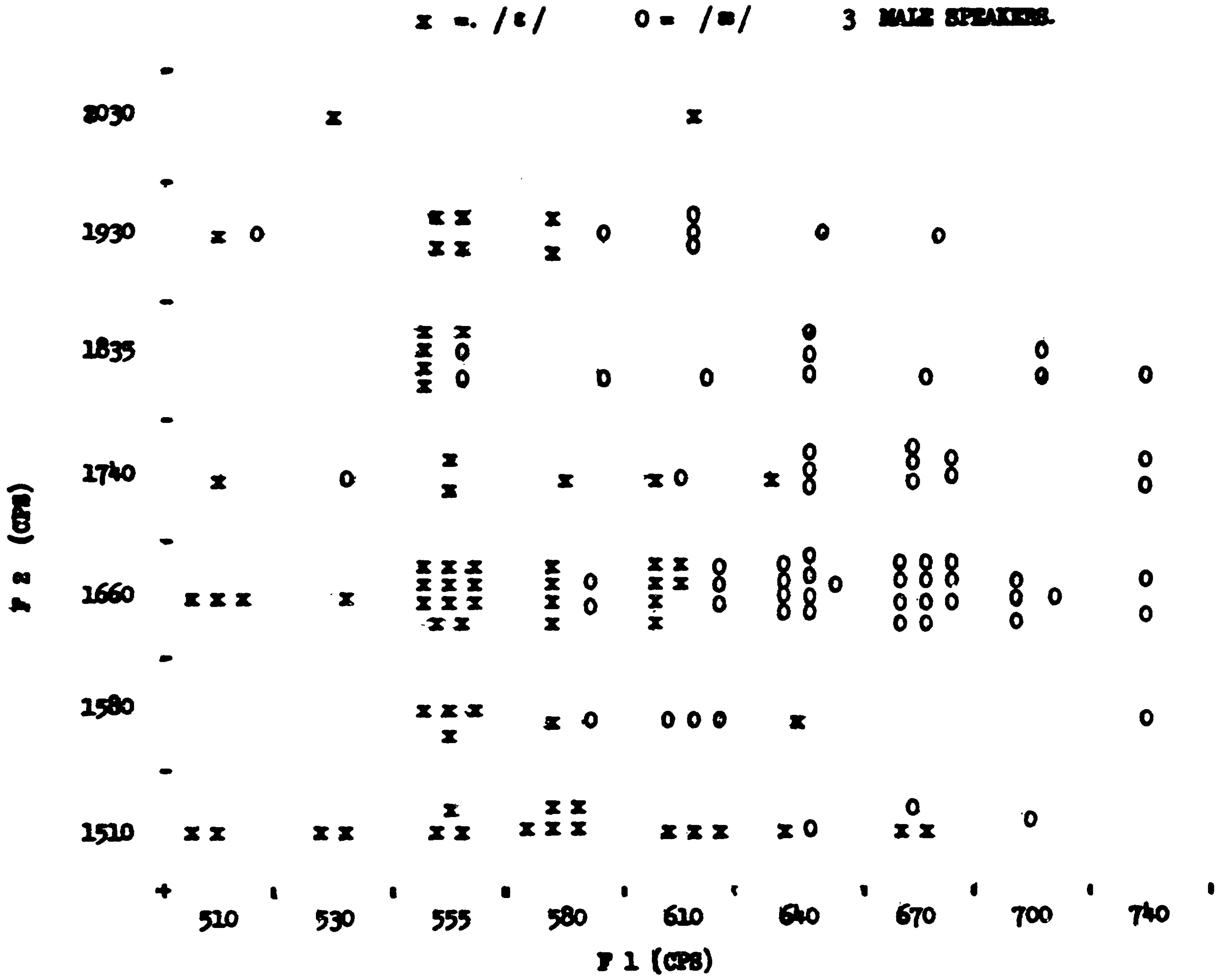

FIG. II-a MEASURED FORMANT FREQUENCIES

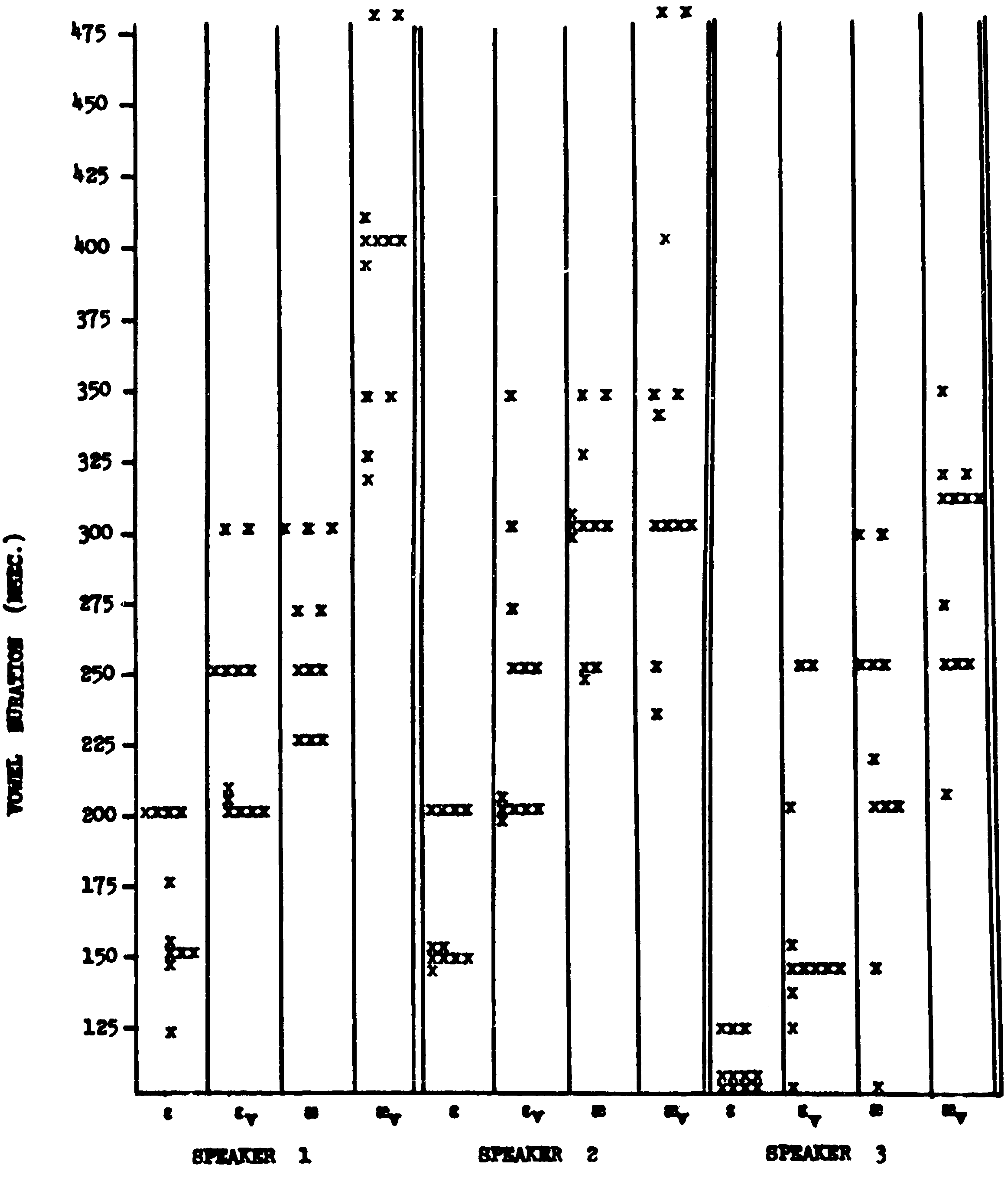

FIG. II-b VOWEL DURATIONS

The last three figures attempt to summarize the formant position, duration, and test data for each of the three speakers separately. Perhaps I am attempting to show too much on one figure, but I feel it is important to see it all at the same time. Color stands for strength or unanimity of listener judgment. Red is for strong, green for medium, and black for weak. The letter represents the speaker's intent (E for /ɛ/, A for /æ/) and the number the original vowel duration in milliseconds. The item is boxed if the listener judgments did not agree with the speaker's intent.

For speaker 1 (Figure II-c), we see a fairly good separation, with perhaps one or two errors, in a simple separation scheme. For speaker 2 (Figure II-d), the separation is not quite as good; but with more reliance on a duration threshold, a good separation scheme could be derived.

Time prevents an adequate discussion of speaker 3's results (Figure II-e). Only seven of his utterances out of 46 were judged to be /æ/. A Yiddish-language background no doubt contributed to this result. Even so, several anolomies in this figure suggest further study of his vowel spectra is in order.

These results can be formulated into a tentative recognition scheme, if so desired. For example, an F1-F2 plane doubtful area could be defined from which the scheme would refer the utterance to a length criterion, based on information as to the speakers' vowel length habits and on information as to the voicing of the following phoneme. I do not suggest rules at this point, however, but only wish to point out that at least a simple strategy becomes possible.

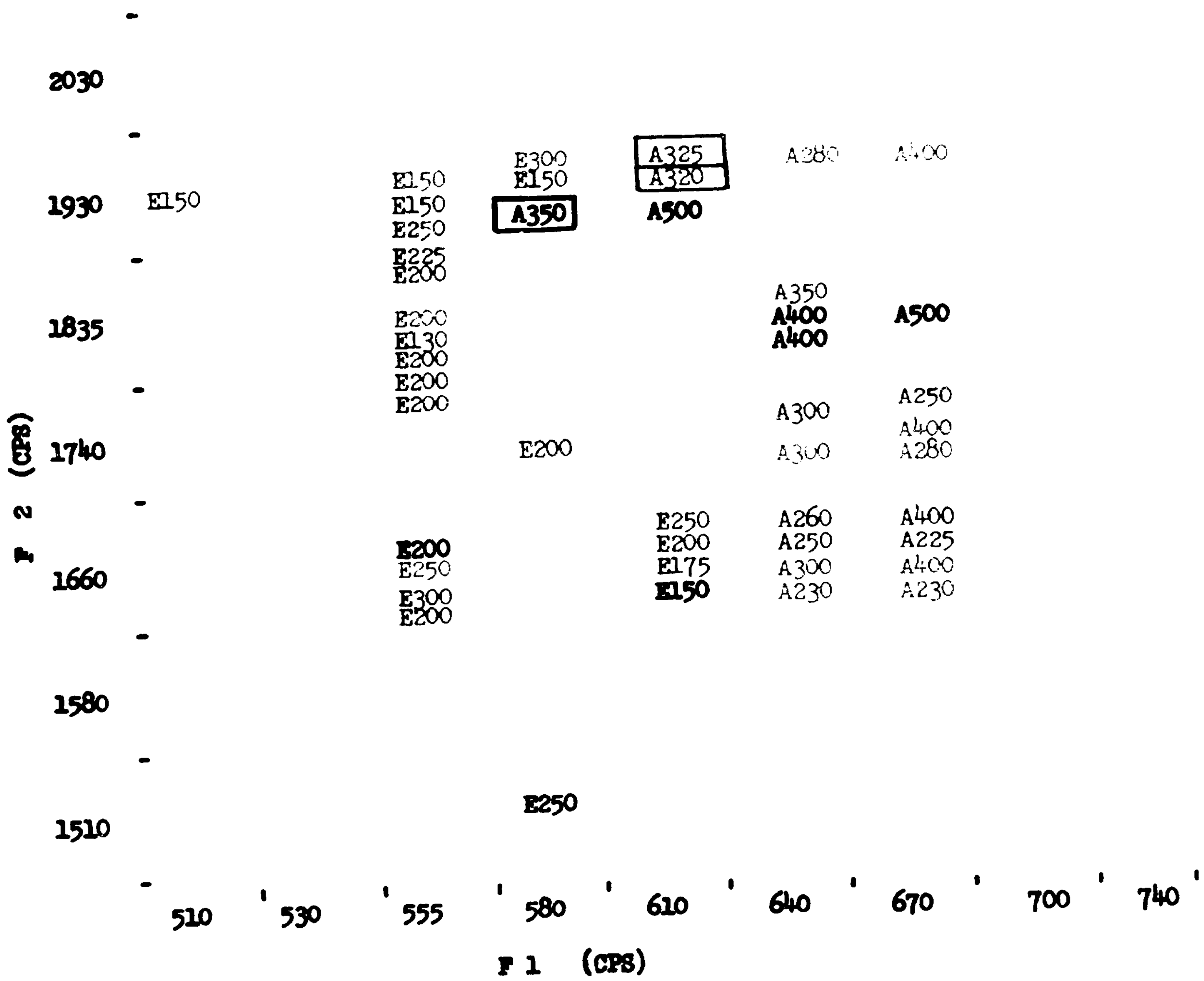

FIG. II-c /ɛ/ - /æ/ SPEAKER 1

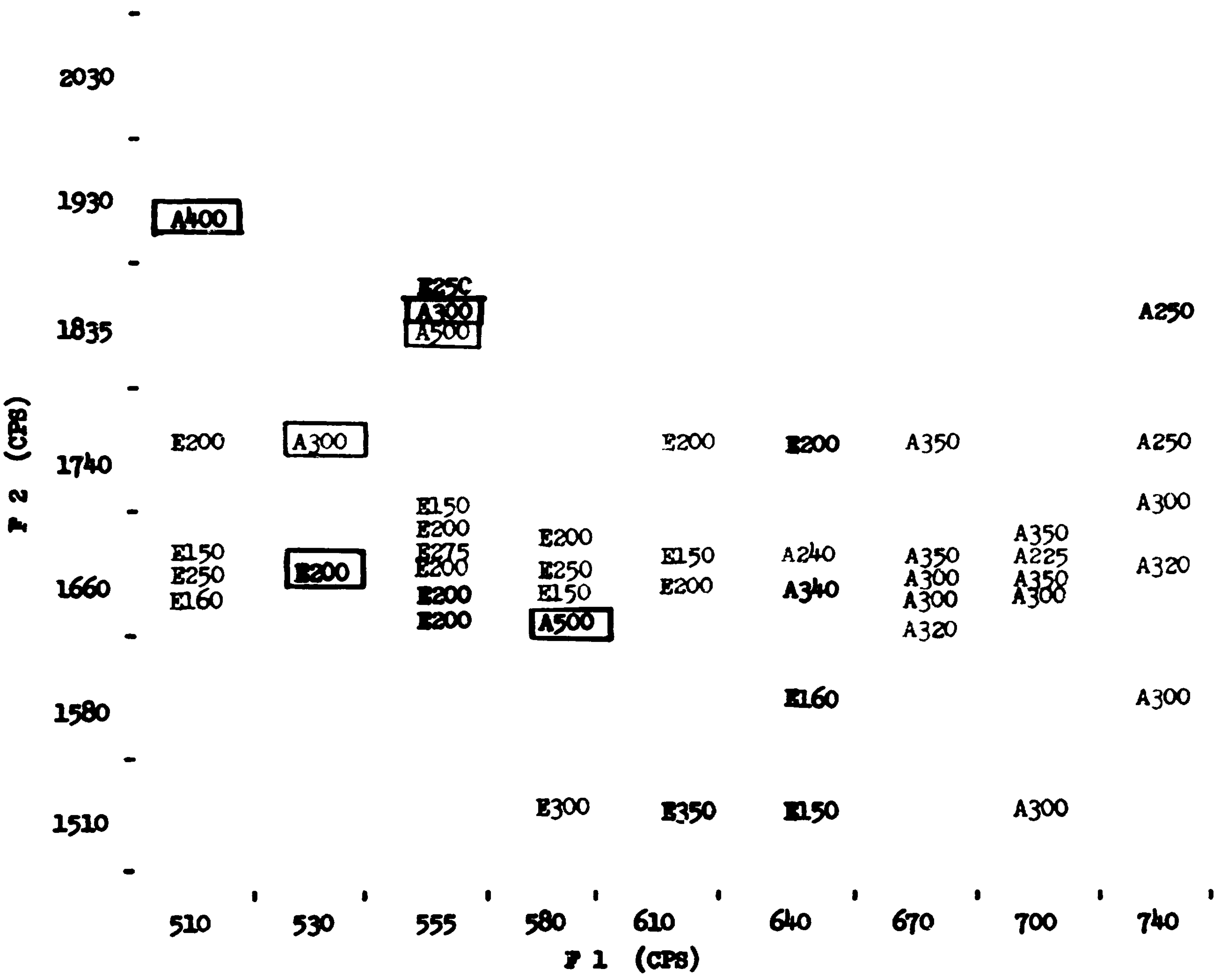

FIG. II-d /ɛ/ - /æ/ SPEAKER 2

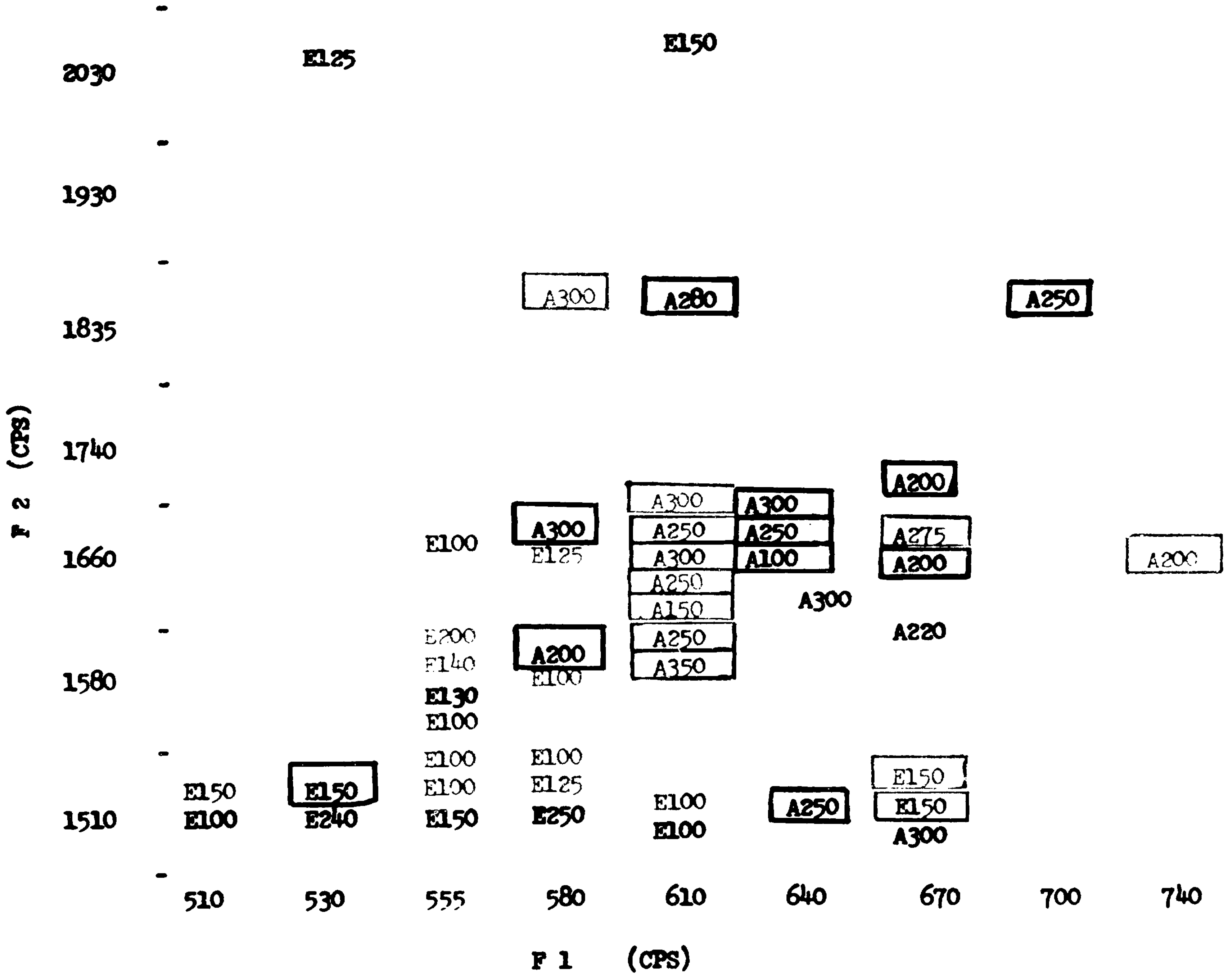

FIG. II-e /ε/ - /æ/ SPEAKER 3

Appendix III

SPEECH DATA PROCESSING EQUIPMENT

After the desired speech utterances were recorded on one channel of a continuous loop of magnetic tape, a series of 60 cps pulses were recorded on the other tape channel to provide sample timing. Each repetition of the utterance loop rotation advanced the position of a telephone-type stepping switch which selected the rectified and smoothed output of one of the 35 band-pass filters. The filters, with typical characteristics as shown in Figure III-b, are listed in Figure III-a along with their center frequencies and bandwidths.

After rectification and smoothing (with time constants proportional to center frequency), these outputs were sampled 60 times a second, coincident with the pulses recorded on the second channel of the magnetic tape. A Datrac analog-to-digital converter produced a ten-bit representation of each point on the speech spectrum which was then fed into a Royal Precision RPC 4000 digital computer. This computer transposed the data into a frequency-time matrix and controlled an IBM card punch.

The output of the card punch thus consisted of a deck of cards on which appeared the amplitude of each filter output during each sample interval. All recognition programs were written for the IBM 7090 computer which uses the IBM cards as its input.

A block diagram of the input system is shown in Figure III-c.

Further details on the data processing equipment may be found in Reference 10.

Filter No.		LHP	UHP	BW
1	286	250	296	46
2	317	302	347	45
3	368	350	396	46
4	428	400	450	50
5	473	450	496	46
6	526	500	552	52
7	585	553	610	57
8	643	611	674	63
9	707	675	743	68
10	780	744	821	77
11	864	825	911	86
12	966	912	1005	93
13	1070	1006	1112	106
14	1157	1112	1230	118
15	1290	1229	1353	124
16	1425	1361	1499	138
17	1560	1501	1656	155
18	1713	1661	1831	170
19	1901	1837	2026	189
20	2087	2030	2238	208
21	2316	2241	2475	234
22	2550	2477	2739	262
23	2814	2734	3025	291
24	3114	3032	3338	306
25	3476	3331	3681	350
26	3831	3692	4074	382
27	4255	4073	4498	425
28	4647	4502	4976	474
29	5148	4984	5489	505
30	5671	5495	6065	570
31	6284	6086	6719	633
32	6951	6716	7425	709
33	7800	7436	8204	768
34	8600	8223	9084	861
35	9500	9076	10,039	963

Figure III-a. Table of Filter Numbers and Frequencies.

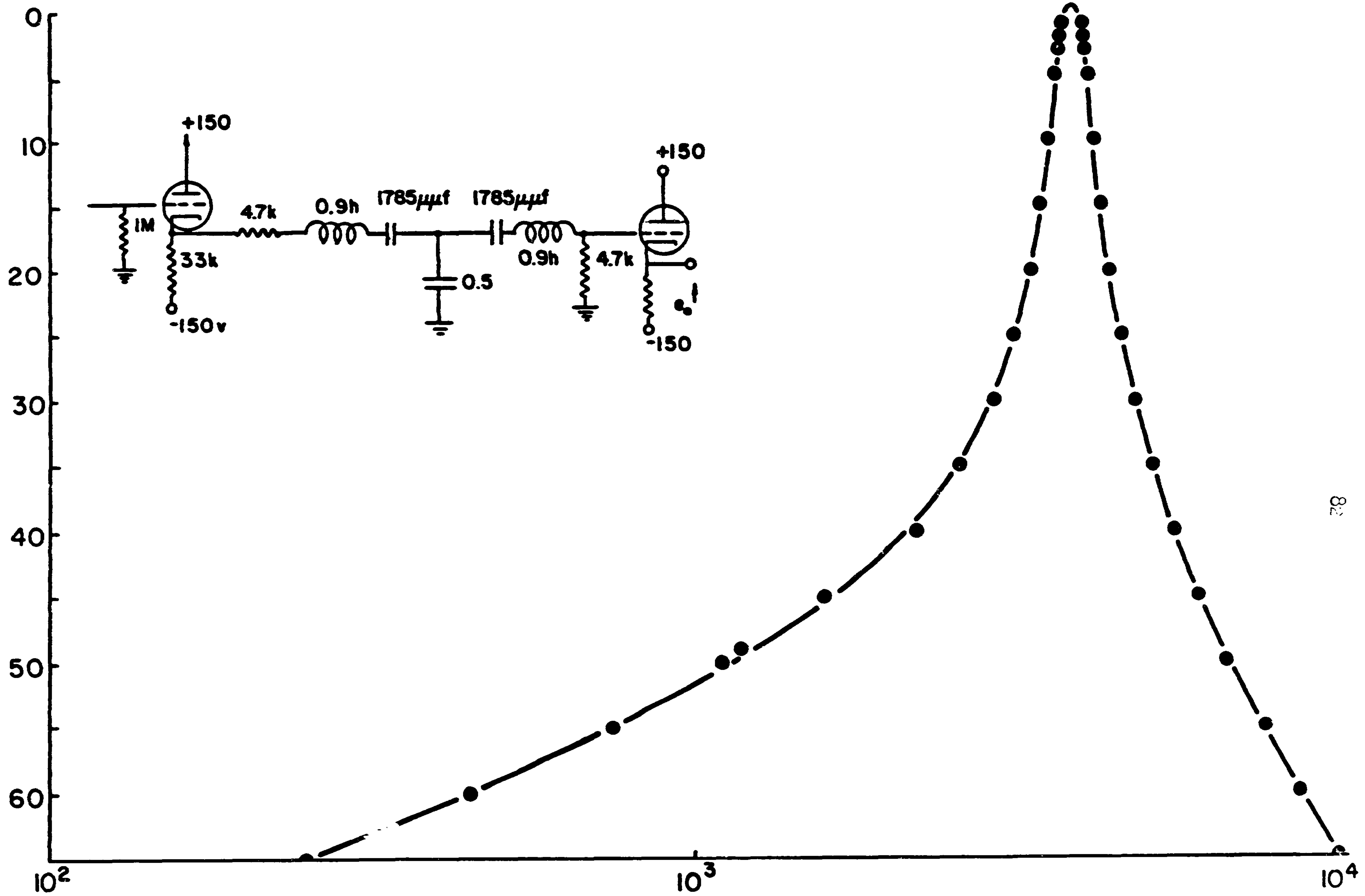

Figure III-b. Typical Filter Characteristics.

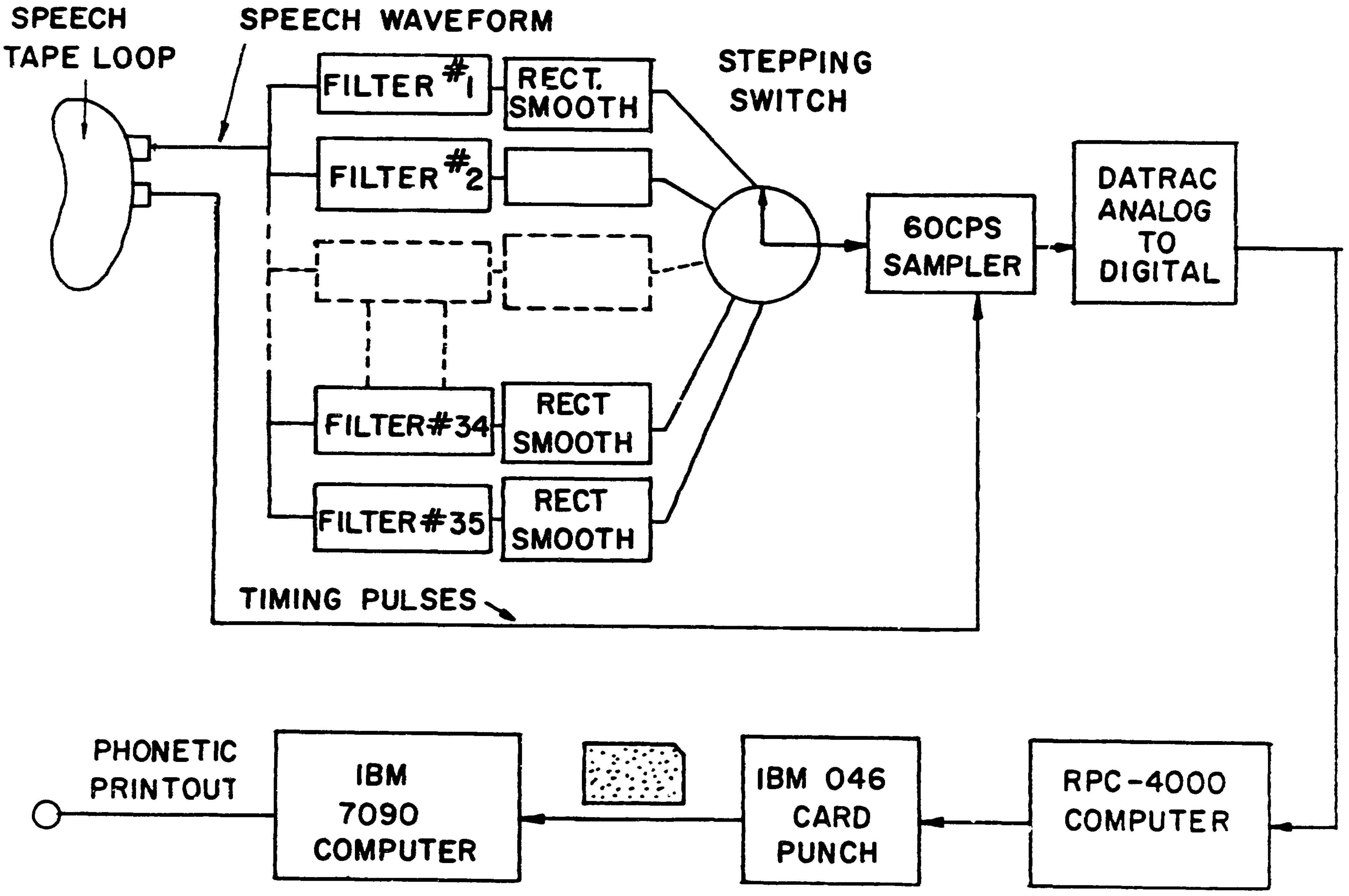

Figure III-c. Block Diagram Speech Processing Equipment.

Appendix IV

LOW-FREQUENCY HARMONIC STRUCTURE

Preliminary study of non-vowel sonorants as a class indicates that many important acoustic cues reside in the very lowest portion of the speech spectrum - that is, below 500 cps. We feel that the importance of the low-frequency region has been underestimated. A program to utilize such spectral information, both to aid in separating nasals, liquids and glides as a class and to distinguish among them, has begun. Immediately apparent is the difficulty in separating peaks in the spectrum (as evidenced from filter output data) caused by voicing components from those caused by poles of the vocal tract transfer function. To help resolve this problem we have adopted the scheme outlined in Fig. IV-a. Basically, an approximation to the voicing spectrum is calculated for each time segment for which filter-output data exists. This spectrum, passed through simulated filters, is subtracted (all numbers in db) from the original data to produce an approximate measure of the transfer function. The sonorants used in this study were in the inner position (C_2) of bisyllabic utterances /h ə C_1 V_1 C_2 V_2 C_3/.

A. Pitch Tracking Routine

Of first importance to a scheme such as that shown in Fig. IV-a is an accurate pitch-tracking program. The one actually evolved was extremely rapid and simple, involving only manipulations of data from the filter bank. Data from it can be used both for intonation or stress studies and as an input to the transfer function calculations.

The fundamental-frequency tracking routine is as follows:

1. A maximum or peak is determined among the low-frequency filter outputs. The filters examined cover a range from 160-320 cps, which is taken to be approximately the range of the second harmonic of male voicing. To cover this range in actual practice with the filter set described in Appendix III, all utterances were played back at double speed.

2. A pattern consisting of the frequency location of the filter whose output is maximum together with the output levels relative to this maximum of the two adjacent filters is used to obtain a closer approximation to the peak frequency location. This interpolation is done (via table look-up procedure) to the nearest 1/8 tone. The precision

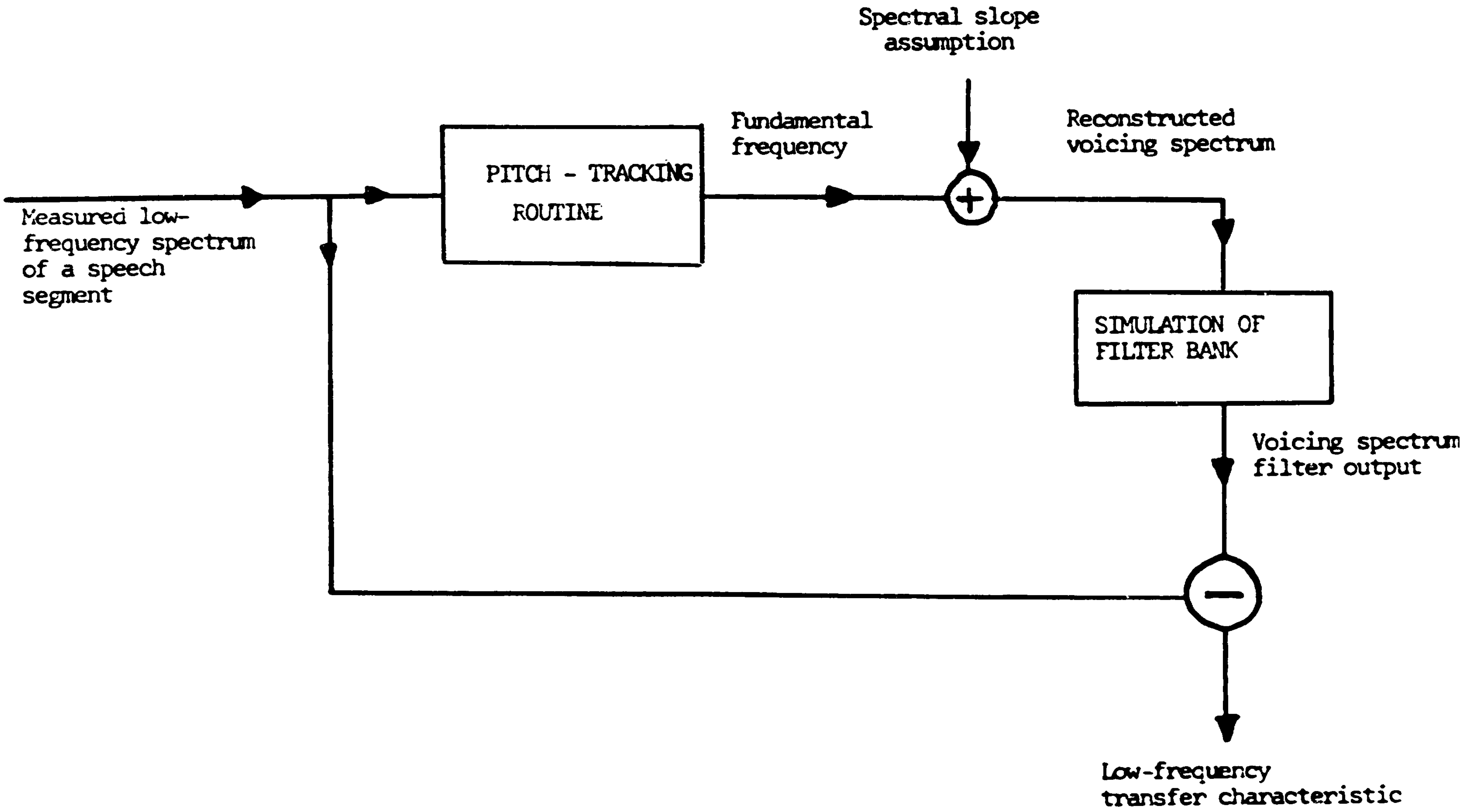

Figure IV-a. Outline of program to obtain sonorant low-frequency transfer characteristics

of locating this peak (presumed to be the second harmonic) at double speed is within ± 10 cps, which is equivalent to ± 2.5 cps for the fundamental frequency.

3. Occasionally the third harmonic will enter this frequency range and cause the spectral peak. It was relatively easy to program a decision as to whether the second or third harmonic had been located by examining the outputs of filters lower in frequency than the maximum. The difference between these patterns is so gross that even the presence of a resonance in or near the second harmonic region will not prevent a reliable decision.

4. An appropriate division of the frequency of the peak by two or by three gives the fundamental frequency, f_o. This output correlates very precisely with oscillographically determined f_o, if $85 \leq f_o \leq 150$ cps.

Figure IV-b shows fundamental frequency tracked by this routine for an entire utterance. A measure of sound intensity (the sum of all filter outputs) is also given.

B. Some Sonorant Characteristics

One of the measurable characteristics of most of the utterances of non-vowel sonorants examined so far is apparent on Fig. IV-b. Concomitant to vocal tract constriction is a shift downward (on the order of 10%) of both intensity and pitch.

Some other acoustic features of sonorants can be put in evidence by carrying through the scheme of Fig. IV-a. Some data are shown here for the cases of /w/, /u/, and for contrast, /a/. See Figs. IV-c, IV-d, and IV-e. In all cases the voicing spectrum was determined from the fundamental frequency calculation plus the assumption of an approximately -9 db/octave spectral slope. This assumption has given good calculated versus measured spectral correspondence on all our data thus far, since only one talker has been used.

Of particular note is the difference between the low-frequency resonance of /w/ and that of /u/. Although the resonant frequencies of /w/ shown in Fig. IV-c and those of /u/ shown in Fig. IV-d are approximately equal, the spectrum of /u/ is more in excess of the voicing spectrum than is that of /w/. Also, the "valley" between the resonances is less deep in /w/. These observations support the statement that consonantal resonances are of

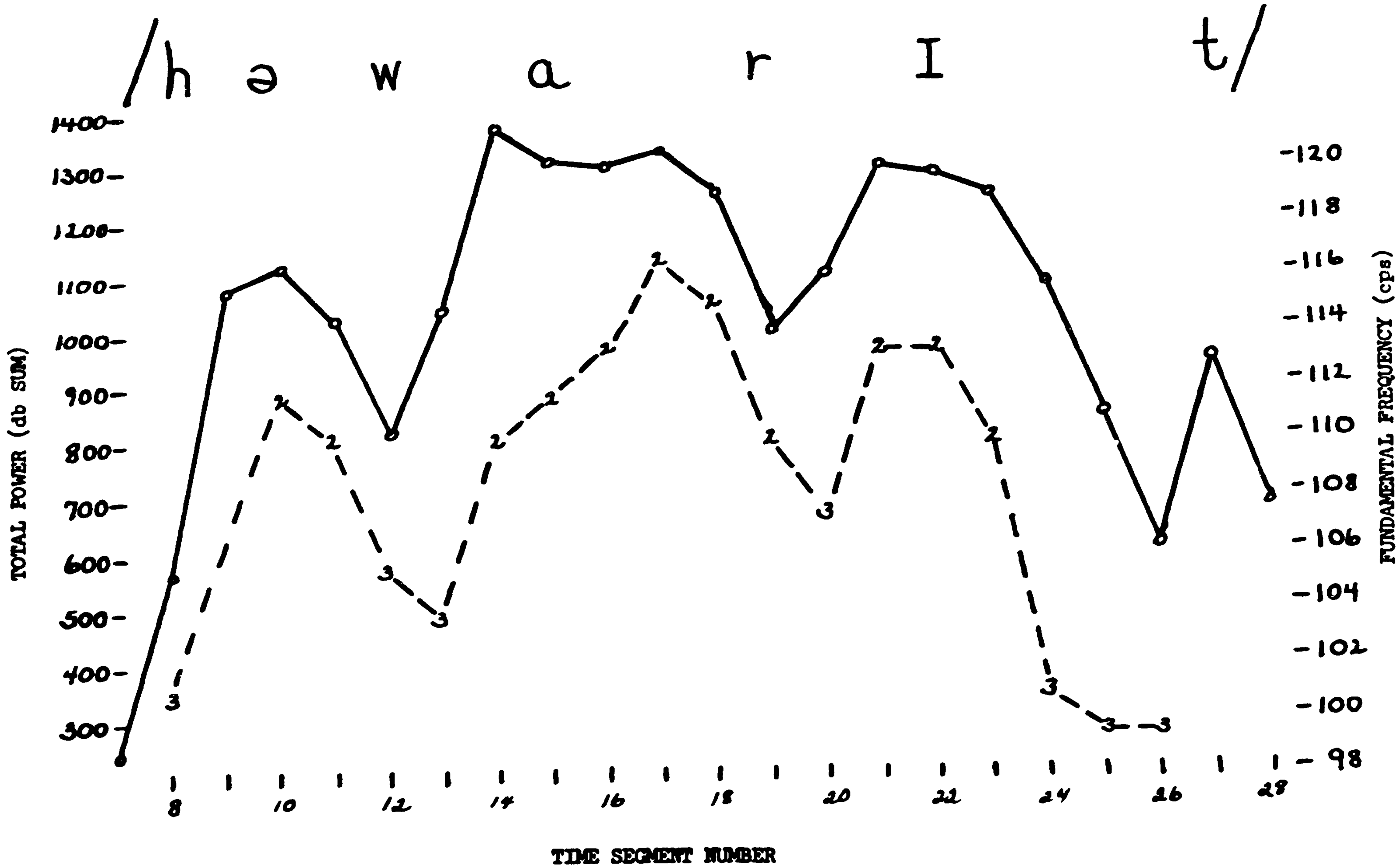

FIG. IV-b. Computer-determined fundamental frequency (slashed line) and total power (solid line) versus time. Real-time scale is 33 msec per segment. Numbers in pitch plot refer to harmonic tracked.

lower "Q" than vowel resonances.

Figure IV-f shows two examples of applying the scheme of Fig. IV-a. Both are /w/ and show relatively broad resonances below 500 cps. The transfer characteristic clearly can be influenced by surrounding vowels in the region from 500-900 cps, but is more consistent, at least for this speaker, outside this range.

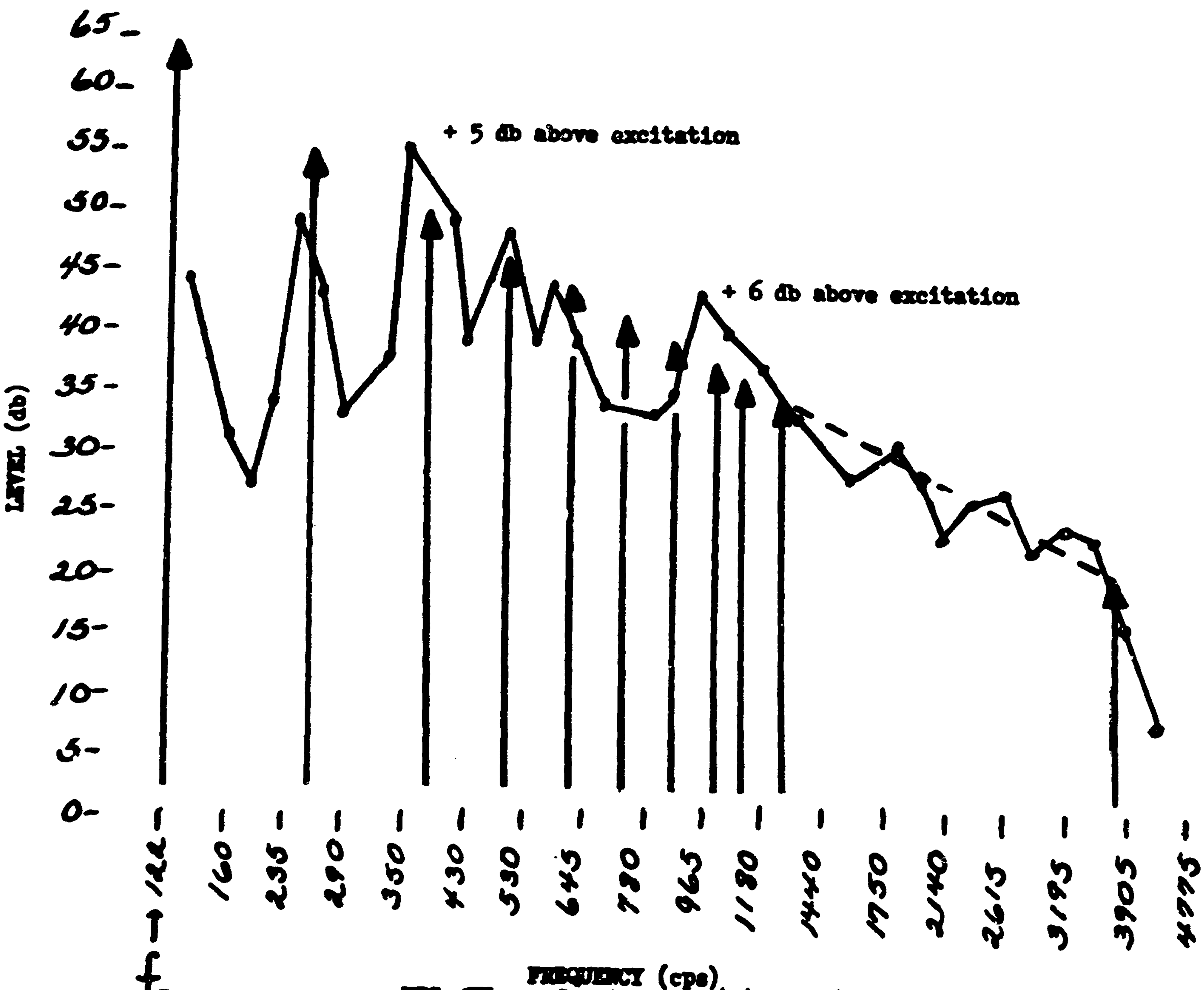

FIG. IV-c. Spectrum of /w/ from /h e m ɛ w a n/ together with voicing spectrum.

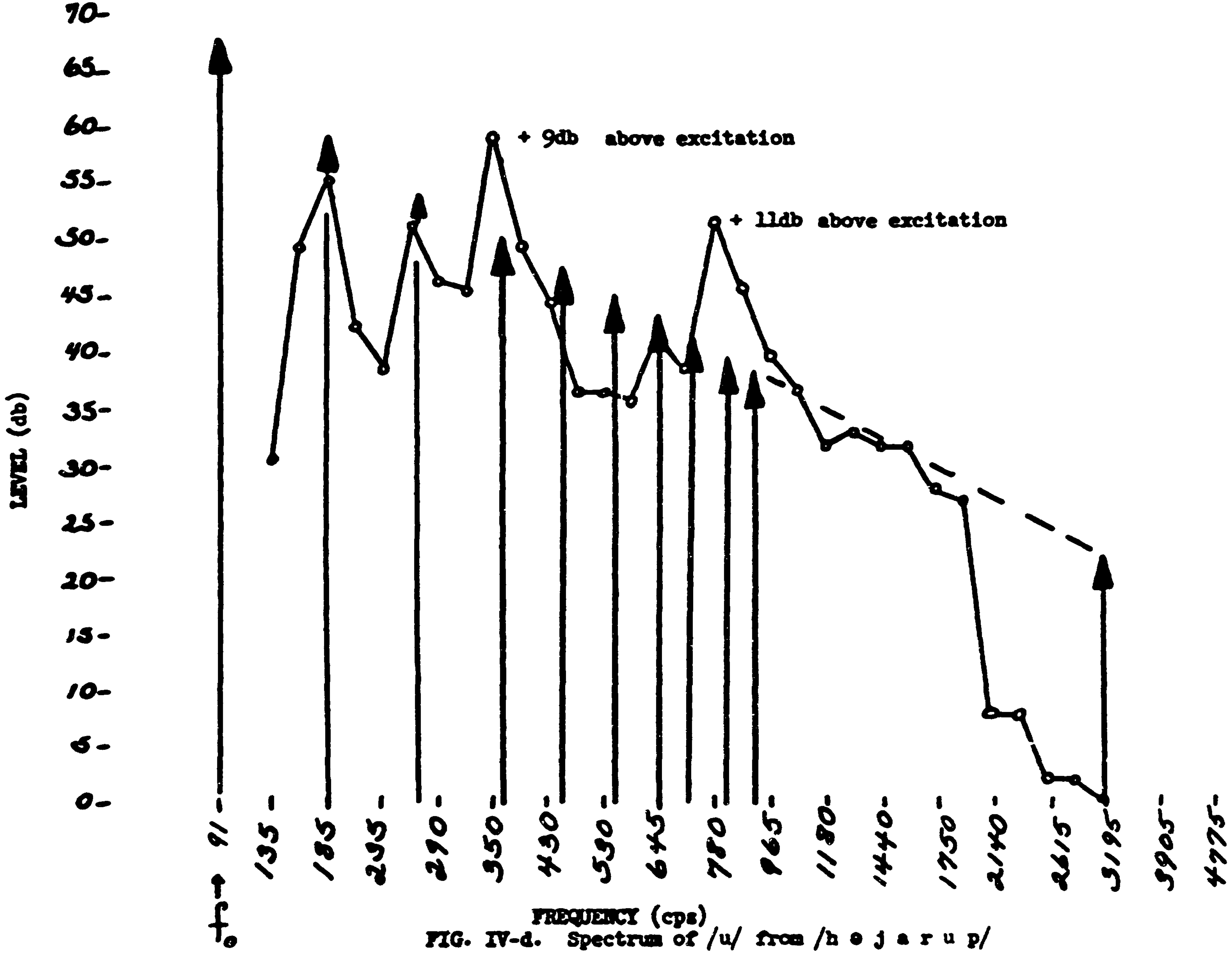

FIG. IV-d. Spectrum of /u/ from /h e j a r u p/ together with voicing spectrum.

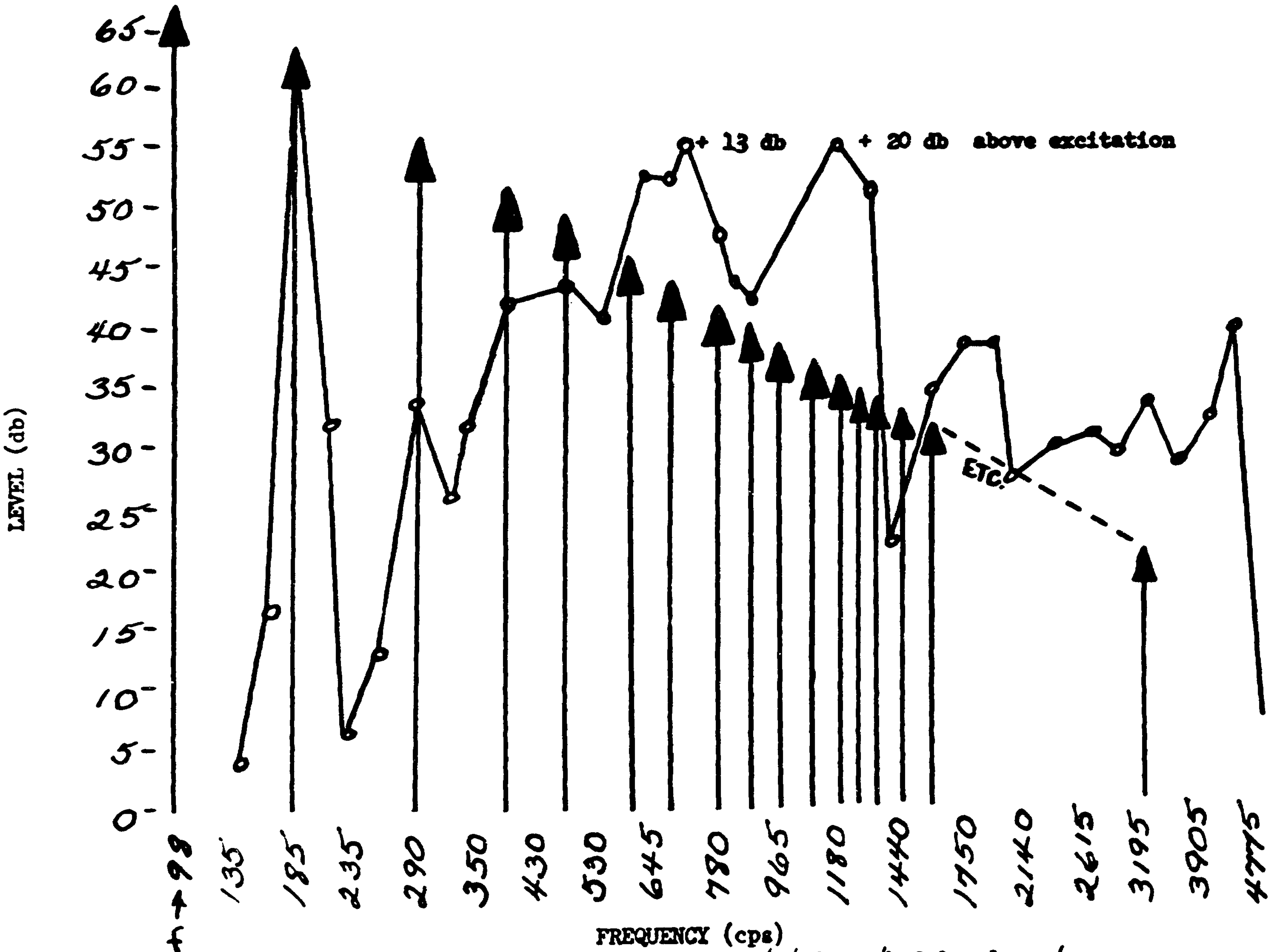

FIG. IV-e. Spectrum of /a/ from /h ə l u l a p/ together with voicing spectrum.

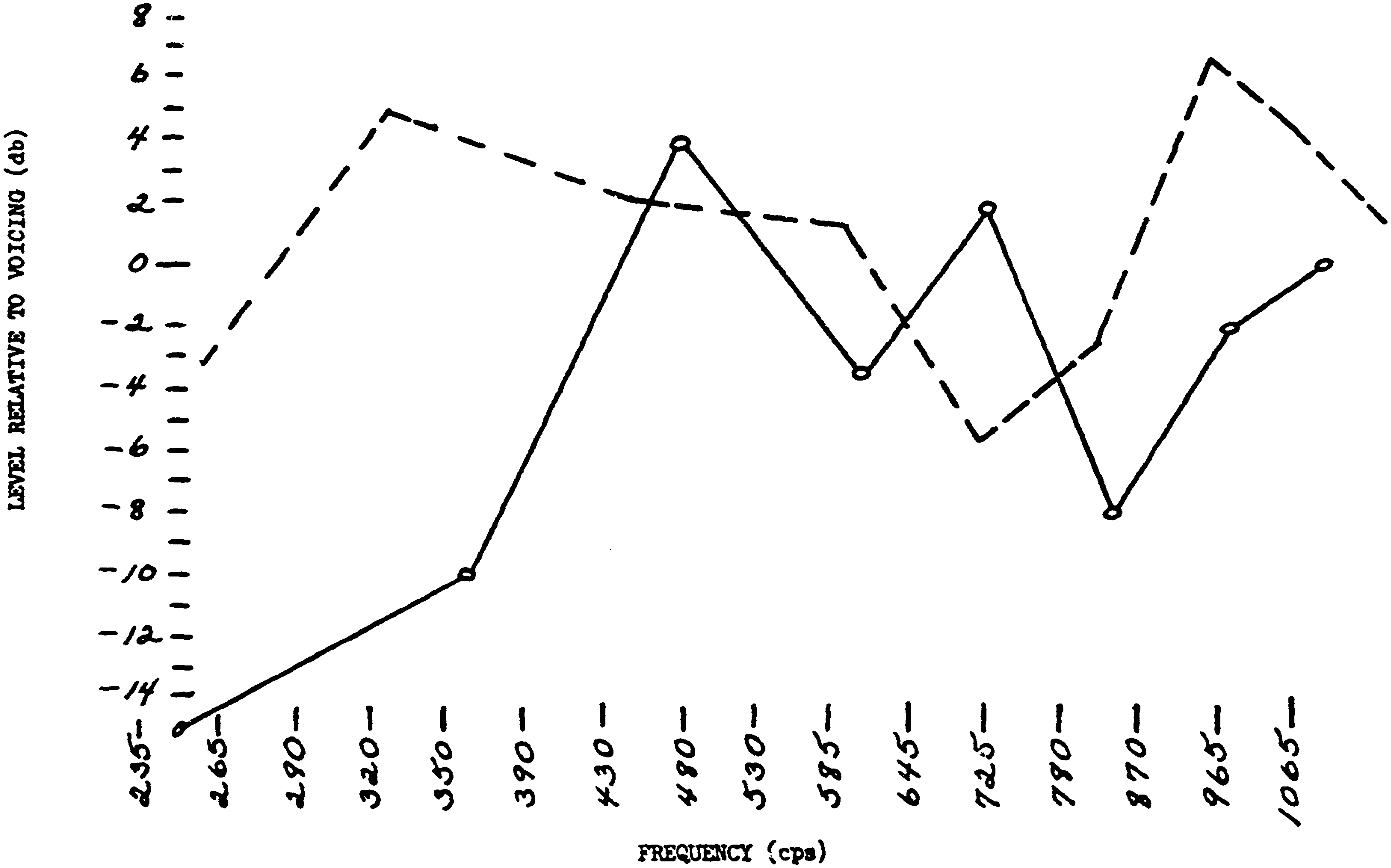

FIG. IV-f. Transfer characteristics of two samples of /w/. From /h ə ʃ a w æ z/ - solid line From /h ə m æ w a n/ - slashed line.

REFERENCES

1. BELL, C. G., FUJISAKI, H., HEINZ, J. M., STEVENS, K. N., and HOUSE, A. S. Reduction of speech spectra by analysis-by-synthesis techniques. J. Acoust. Soc. Am., 33: 1725-1736, 1961.

2. CHERRY, E. C., HALLE, M., and JAKOBSON, R. Toward the logical description of languages in their phonemic aspect. Language, 29: 34-46, 1953.

3. FANT, C. G. M. Acoustic analysis and synthesis of speech with applications to Swedish. Ericsson Technics, No. 1, pp. 3-108, 1959.

4. FANT, C. G. M. Acoustic Theory of Speech Production. 's-Gravenhage, Mouton and Co., 1960.

5. FRY, D. B. Perception and recognition in speech. For Roman Jakobson. 's-Gravenhage, Mouton and Co., p. 169, 1956.

6. HALLE, M. The strategy of phonemics. Word, 10: 197-209, 1954.

7. HALLE, M. Why and how do we study the sounds of speech? Monograph No. 7, Georgetown University Monograph Series on Language and Linguistics, Sept. 1954.

8. HALLE, M. In defense of the number two. Studies Presented to J. Whatmough. 's-Gravenhage, Mouton and Co., pp. 65-72, 1957.

9. HALLE, M., and STEVENS, K. Speech recognition: a model and a program for research. IRE Trans. on Information Theory, Vol. IT-8, No. 2, pp. 155-159, Feb. 1962.

10. HEMDAL, J. F. The Application of Distinctive Features to the Primary Recognition of Speech. Ph.D. Thesis, Purdue University School of Electrical Engineering, 1964.

11. HOUSE, A. S. On vowel duration in English. J. Acoust. Soc. Am., 33: 1174-1178, 1961.

12. HUGHES, G. W. The Recognition of Speech by Machine, Tech. Report No. 395, Res. Lab. Elect., M. I. T., Cambridge, Mass., 1961.

13. HUGHES, G. W. and HALLE, M. On the recognition of speech by machine. Proceedings of the International Conference on Information Processing, Paris, Oldenburg, 1959.

14. JAKOBSON, R., FANT, C. G. M., and HALLE, M. Preliminaries to Speech Analysis, Tech. Report No. 13, Acoust. Lab., M. I. T., Cambridge, Mass., 1952.

15. JAKOBSON, R., and HALLE, M. Fundamentals of Language, 's-Gravenhage, Mouton and Co., 1956.

16. LADEFOGED, P., and BROADBENT, D. E., Information conveyed by vowels. J Acoust. Soc. Am., 29: 98-104, 1957.

17. LAWRENCE, W. The synthesis of speech from signals which have a low information rate. Proc. of the 1952 Symposium on the Applications of Communication Theory, ed. W. Jackson, London, pp. 460-569, 1953.

18. LEHISTE, I., and PETERSON, G. E. Transitions, glides and diphthongs. *J. Acoust. Soc. Am.*, *33*: 268, 1961.

19. LIEBERMAN, P. Some effects of semantic and grammatical context on the production and perception of speech. *Lang. and Speech*, *6*: 172-187, 1963.

20. LIEBERMAN, P. Intonation and the syntatic processing of speech. *Proc. of Symposium on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, 1965.

21. LISKER, L. Minimal cues for separating /w, r, l, j/ in intervocalic position. *Word*, *13*: 256-267, 1957.

22. MOL, H., and UHLENBECK, E. M. Hearing and the concept of the phoneme. *Lingua*, *8*: 161-185, 1959.

23. PETERSON, G. E., and BARNEY, H. L. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, *24*: 175-184, 1952.

24. POZA, F. *Formant Tracking by Digital Computation*. M. S. Thesis, M. I. T., Dept. of Elect. Eng., Sept 1959.

25. STEVENS, K. N., and HALLE, M. Remarks on analysis-by-synthesis and distinctive features. *Proc. of Symposium on Models for the Perception of Speech and Visual Form*, AFCRL, Boston, 1965.

26. WIIK, K. Phoneme boundaries of Finnish vowels. *Proc. of the IVth International Congress of Phonetic Sciences*, Helsinki, 1961.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Purdue Research Foundation
Lafayette, Indiana

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE

Speech Analysis

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Scientific Report Period covered: January 1962-December 1964

Approv d:
12 Nov 65

5. AUTHOR(S) *(Last name, first name, initial)*
Hughes, George W. and Hemdal, John F.

6. REPORT DATE
July 1, 1965

7a. TOTAL NO. OF PAGES
100

7b. NO. OF REFS
26

8a. CONTRACT OR GRANT NO.
AF19(628)-305

b. PROJECT AND TASK NO.
5628-02

c. DOD ELEMENT
61445014

d. DOD SUBELEMENT
681305

9a. ORIGINATOR'S REPORT NUMBER(S)
TR EE65-9

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
AFCRL-65-681

10. AVAILABILITY/LIMITATION NOTICES Qualified requestors may obtain copies of this report from DDC. Other persons or organizations should apply to the Clearinghouse for Federal Scientific and Technical Information (CFSTI), Sills Building, 5285 Port Royal Road, Springfield, Virginia 22151.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Hq. AFCRL, OAR (CRB)
United States Air Force
L.G.Hanscom Field,Bedford,Mass.

13. ABSTRACT

The limitations of speech recognition procedures which depend solely on acoustic data are discussed. One such "primary recognition" scheme, based on phoneme classification by tracking the acoustic correlates of a set of distinctive features, is presented. Programmed on a digital computer, these logical operations on digitalized spectra of 17-msec samples of speech were tested on some 300 nonsense utterances from two talkers. A priori information about individual talker characteristics is incorporated into the logic (single-speaker approach). Comparison of machine performance was made with both the intent of the speaker and with the judgments of listeners. Listeners were presented with the same acoustic stimuli that were machine processed. Some perceptual tests were run on short vowel segments excised from nonsense syllables.

Detailed quantitative results are presented only for vowels. They show that man and machine agree about 90% of the time on vowel judgments under these conditions of minimal contextual information. Clear feature boundaries are shown on the F1-F2 plane for the (stressed) vowel utterances. Although these boundaries are not generally valid for more than one voice, simple translations of them may suffice to obtain usable vowel separation for many talkers.

Unclassified

Security Classification

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Speech recognition						
Phoneme classification						
Acoustical phonetics						
Digital computer						
Speech perception						
Speech spectra						
Distinctive features						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
_______________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
_______________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
_______________."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS), (S), (C), or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

Unclassified
Security Classification